IMPACT OF ACTION-OBJECT CONGRUENCY ON THE INTEGRATION OF AUDITORY AND VISUAL STIMULI IN EXTENDED REALITY

A Dissertation
Presented to
The Academic Faculty

by

Keenan R. May

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Psychology

Georgia Institute of Technology

May 2020

IMPACT OF ACTION-OBJECT CONGRUENCY ON THE INTEGRATION OF AUDITORY AND VISUAL STIMULI IN EXTENDED REALITY

Approved by:

Dr. Bruce Walker, Advisor
School of Psychology
Georgia Institute of Technology

Dr. Thackery Brown
School of Psychology
Georgia Institute of Technology

Dr. Jamie Gorman
School of Psychology
Georgia Institute of Technology

Dr. Maribeth Gandy
Interactive Media Technology Center
Georgia Institute of Technology

Dr. Richard Catrambone
School of Psychology
Georgia Institute of Technology

Date Approved: [November 20, 2019]

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Bruce Walker, for providing the guidance and encouragement I needed, the School of Psychology for providing an environment in which to grow, and the members of the Sonification Lab, past and present, for their unwavering support, wisdom and friendship.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

XR	Extended Reality
VE	Virtual Environment
HMD	Head Mounted Display
MSI	Multisensory Integration
STS	Superior Temporal Sulcus
STG	Superior Temporal Gyrus
MTL	Mediotemporal Lobe
TOJ	Temporal Order Judgment
JND	Just-Noticeable Difference
SOA	Stimulus Onset Asynchrony
HRTF	Head-Related Transfer Function
RSI	Response-Stimulus Interval
SSQ	Simulator Sickness Questionnaire

SUMMARY

Extended Reality (XR)[1] systems are currently of interest to both academic and commercial communities. XR systems may involve interacting with many objects in three-dimensional space. The usability of such systems could be improved by playing sounds that are perceptually integrated with visual representations of objects. In the *multisensory integration* process, humans take into account various types of crossmodal congruency to determine whether auditory and visual stimuli should be bound into unified percepts. In XR environments, spatial and temporal congruency may be unreliable. As such, the present research expands on *associative congruency*, which refers to content congruency effects that are acquired via perceptual learning in response to exposure to co-occurrent stimuli or features. A new type of associative congruency is proposed called *action-object congruency*. Research in ecological sound perception has identified a number of features of *objects* and *actions* that humans can discern based on the sounds produced by sound-producing events. Since humans can infer such information through sound, this information should also inform the integration of auditory and visual stimuli. When perceiving a realistic depiction of a sound-producing event such as a strike, scrape or rub, integration should be more likely to occur if a concurrently-presented sound is congruent with the objects and action that are seen. These effects should occur even if the visual objects and the sound are novel and unrecognizable, as long as relevant features can be ascertained visually and via sound. To evaluate this, the temporal and spatial ventriloquism illusions were utilized to assess the impact of action congruency and object congruency on

[1] A term that encompasses Virtual Reality, Mixed Reality, and Augmented Reality

multisensory integration. Visual depictions of interacting objects were displayed in virtual reality, and congruent or incongruent sounds were played over speakers. In two types of trials, participants either localized the sounds via pointing, or judged whether the sounds and visual events were simultaneous. Action-object congruent visual and auditory pairings led to greater localization biasing and higher rates of perceived simultaneity, reflecting stronger integration of stimuli. Action and object congruency were both impactful, but action congruency had a larger effect. The effects of action and object congruency were additive, providing support for the linear summation model of congruency type combination. These results suggest that action-object congruency can be used to better understand how humans conduct multisensory integration as well as to improve MSI in future XR environments.

CHAPTER 1. INTRODUCTION

Developments in head-mounted display (HMDs) have led to the proliferation of virtual reality technologies, in which virtual objects replace reality, or augmented and mixed reality experiences in which virtual objects are blended with reality. A persistent issue faced by all three types of systems (collectively, "extended reality" or XR) is the "out-of-view problem," which refers to the tendency for users to lose track of objects outside of the often-narrow field of view of the HMD. Compounding the problem is the fact that human vision itself has a limited field of view, relative to the possible locations of objects within room-scale (or larger) XR environments. In such environments, rather than virtual objects being constrained to a two-dimensional display within a user's field of view, objects can be placed arbitrarily in a three-dimensional space. In addition to being out of view due to head movement, objects can move, become occluded by virtual or real objects, or be left behind in another room. Existing user interface principles do not sufficiently support interacting with XR environments comprised of such objects.

An early proposal for addressing this issue has been to add visual cues to the location or status of out-of-view virtual objects, in conjunction with basic auditory alerts (Salter et al., 2016). Although potentially helpful, indicators such as these only provide *information* about the direction in which a user should conduct a visual search.

1.1 Facilitating Multisensory Integration for XR Usability

If sounds could be perceptually integrated with of-interest virtual visual objects in the periphery or entirely out of view, this could provide greater benefits in terms of

facilitating correct perception of XR environments. Rather than merely acting as a cue to help direct visual search, integrated audio is be perceived as being *part* of crossmodal objects (Spence, 2011). Multisensory integration (MSI) refers more broadly to the set of processes through which stimuli in different modalities are considered together. In this paper, a specific aspect of MSI called *crossmodal binding* is of primary interest. Crossmodal binding is defined as the attribution of stimuli in different sensory modalities to the same underlying *crossmodal object* (Driver & Spence, 1998). Bizley, Maddox, and Lee (2016, p. 3) defined crossmodal objects as "perceptual construct(s) which occur when a constellation of stimulus features are bound within the brain." Importantly, crossmodal binding is a *perceptual* process, rather than an artefact of human language or decision-making (O'Leary & Rhodes, 1984; Spence, Sanabria & Soto-Faraco 2007; Bizley, Jones, & Town 2016). Scott (2005) suggested that the majority of auditory information in the natural world is in fact processed as status updates from persistent crossmodal objects, rather than as isolated auditory signals (i.e., "sounds"). Crossmodal objects may be initially formed using information from one modality, but can subsequently be updated using information from another modality that is bound to the object (Loomis, Lippa, Klatzky, & Golledge, 2002; Jordan, Clark, & Mitroff, 2010).

There are a variety of benefits to inducing the perception that XR objects are truly multisensory. Multisensory objects are associated with faster response times (Stein & Stanford, 2008). Multisensory objects also tend to be easier than unisensory objects to recognize when visual representations are degraded, and have been recommended for complex display environments that are susceptible to visual noise (Siebold, 2009). Multisensory objects also tend to be more learnable (Fifer, Barutchu, Shivdasani, &

Crewther, 2013). Persons with sensory impairments can benefit especially from effective multisensory virtual displays, through compensatory utilization of alternate modalities when certainty in one modality is low (Laurienti et al., 2003).

The benefit of primary interest to the present discussion is that fact that, in complex environments where many objects may need to be monitored, inducing the perception of objects as being multisensory can facilitate the use of auditory information to provide direct updates to bound crossmodal objects (Spence, Ngo, Lee, & Tan, 2010). Research in aviation in particular has supported the utility of utilizing multisensory objects to increase object tracking performance in complex environments in which visibility may be limited. For example, Bronkhorst, Veltman, and Van Breda (1996) had participants carry out an in-flight following task in which the lead airplane was represented through spatialized audio while it was obscured by clouds. This method was effective at reducing visual search times for the lead airplane. In a similar study, Nelson et al. (1998) found that the use of spatial audio cues led to decreased visual search time for out-of-view objects. There are also performance benefits in applied settings in which objects are entirely visible, but are presented alongside bound auditory display elements in order to assist with perception in complex display environments. For example, Ferris and Sarter (2008) observed improvements to target acquisition speed when co-located auditory and visual display elements were used in a realistic multi-object vehicle monitoring and command task, compared to visual-only displays.

Importantly, bindings of auditory and visual stimuli to crossmodal objects persist over time, and can be utilized after their initial creation, rather than existing only in a brief moment of perception. A set of studies have indicated that recently experienced bindings

persist and decay in working memory, can be actively rehearsed, and lead to increased spatiotemporal tolerances even after a retention interval.

Zmigrod and Hommel (2010) found that crossmodal bindings persist for a period of time, and decay after several seconds. The authors evaluated the manner in which the facilitative effect of an object being multisensory changed alongside increasing response-stimulus intervals (RSIs). At the shortest RSI of 500ms, facilitative effects were largest, and reflected typical speeded classification results. However, these benefits declined as RSIs increased, and appeared to reach asymptote after around three seconds. There is also evidence that this decay can be slowed by active rehearsal, which can in turn be disrupted. Gao et al. (2017) had participants view a series of visual and auditory objects, and later respond with which sound was associated with which visual object. Participants were initially allowed to rehearse these crossmodal objects, which the authors asserted took place using an amodal episodic buffer (Baddeley, Allen, & Hitch, 2010). When participants completed a rehearsal-interferent task in which they were required to utilize object-based attention during the retention interval, performance decreased on the binding memory task, in a continuous manner that was in alignment with typical working memory decay.

In a related study, Piemo, Caria, and Castiello (2006) found evidence that stimuli from multiple modalities could remain integrated for a "long period of time." The authors had participants in a VE locate a visual object that was out-of-view. Their search was guided either by light or sound emitted by the object. Participants were able to locate the object faster when both light and sound were emitted. If either the visual or the auditory cue was spatially or temporally incongruent, this facilitative effect went away. Importantly, the spatial audio cues continued to assist object search even when that search took some

time to carry out, which the authors attributed to those bindings persisting in working memory throughout the subsequent visual search task.

Even if bindings have decayed out of working memory, crossmodal perceptual learning occurs rapidly and flexibly, and can facilitate later crossmodal binding of stimuli (Van Wanrooij, Bremen, & John Van Opstal, 2010; Piazza, Denison, and Silver, 2018, see section 1.3.2.1).

Thus, in addition to multisensory objects being more easily tracked and processed if the objects are within a person's field of view, previously-created bindings could be leveraged to create direct perceptual updates to entirely out-of-view XR objects, as long as a successful instance of binding can be induced initially. The following section discusses common methods for inducing binding in XR, and why these may be inadequate to reliably achieve this goal.

1.2 Spatiotemporal Congruency and XR

XR systems tend to rely in large part on the use of spatialized audio and synchronous presentation, or *spatiotemporal congruency*, to facilitate crossmodal binding. Although maintaining spatiotemporal congruency is important, estimates of both the time of occurrence and the spatial position of visual depictions of objects as well as sounds are often unreliable in XR, suggesting that spatiotemporal congruency alone may be insufficient to reliably induce binding.

1.2.1 Spatiotemporal Congruency Overview

Temporal congruency refers to whether stimuli appeared to occur at the same time, and spatial congruency refers to whether stimuli in different modalities are localized to the same point in space). Spatiotemporal congruency has been well-studied, in large in the form of research into temporal and spatial "binding windows," also referred to as tolerances. This body of research has found that the likelihood of a participant forming a unified percept tends to decrease in a predictable fashion as signals differ more so in space and time, with the size of these spatial and temporal binding windows being influenced by other forms of crossmodal congruency.

The presence of these binding windows means that stimuli from different modalities may be bound together even if they are not completely spatiotemporally congruent, which can be observed behaviorally in the form of subsequent biasing of location and/or time estimates. The "ventriloquism effect" has been the prevalent research paradigm used to study spatiotemporal congruency, as well as MSI in general. *Spatial ventriloquism* occurs when the location estimate of a stimulus in one modality biases the perceived location of a bound stimulus in another modality. Similarly, *temporal ventriloquism* refers to cases when the timing of a stimulus on one modality biases the perceived timing of a bound stimulus in another modality.

Temporal order judgment (TOJ) tasks, simultaneity judgment tasks, and spatial ventriloquism tasks are common paradigms that have been used to investigate crossmodal binding via temporal and spatial ventriloquism.

In TOJ tasks, participants are tasked with responding with which of two stimuli appeared first. Often, sounds are presented before the first visual stimulus and after the second stimulus, which can influence the time that each target stimuli is perceived (i.e., temporal ventriloquism), which can in turn increase or decrease TOJ performance depending on the degree of temporal offset (as well as the congruency of the sounds and visuals). For example, Morein-Zamir, Soto-Faraco, and Kingstone (2003) utilized a TOJ paradigm, and found that playing a sound just before the first visual stimulus or just after the second led to finer just-noticeable differences (JNDs). Additional experiments ruled out the possibility of the sounds simply being alerts, by demonstrating that playing the sounds *between* the two light flashes led to a *decrease* in TOJ performance. Temporal ventriloquism effects were observed up to around 225 ms SOA.

In simultaneity judgment tasks, stimulus presentation is similar, but participants are asked whether stimuli occurred at the same time. These stimuli are often located in slightly different locations, which allows for simultaneous presentation. Hirsh and Sherrick Jr (1961), using an early instance of a simultaneity judgment task, found that their participants became sensitive to audiovisual asynchrony at 20 ms SOA.

Finally, in spatial ventriloquism tasks, participants are asked to point toward the location of sounds, in the presence of visual objects that may bias localization of the sounds if they are bound (e.g., Bertelson & Aschersleben, 1998; Bruns and Röder, 2019).

Using the aforementioned methods, temporal and spatial binding windows have been found to interact. When stimuli are moved closer together in space, temporal binding windows widen; conversely, increases in temporal congruency lead to widened spatial

tolerances, as long as stimuli were not extremely far away in time or space (Slutsky and Recanzone, 2001; Zampini, Guest, Shore, & Spence, 2005). However, if spatial and temporal congruency are both high, slight perturbations to one or the other may not have an observable effect on MSI. Vroomen and Keetels (2006) found that varying the preceding sound in a TOJ task by several degrees was not enough to disrupt the TOJ performance facilitation effect. There are also significant individual differences in spatial and temporal binding windows. One salient example is that of musicians, who tend to have temporal binding windows that are around 1/3 the size of those of non-musicians (Bidelman, 2016).

Spatiotemporal congruency is associated with the activity of single multisensory neurons in the superior colliculus (SC, Stein & Stanford; 2008). These neurons are responsive to temporally synchronous and spatially co-located auditory and visual stimuli, but not to those same auditory or visual stimuli presented on their own. They have receptive fields that are spatially tuned and require near-simultaneous inputs from multiple unisensory areas. Activation spikes tend to be observed around 100 ms post stimulus onset, indicating an early process. Importantly, multisensory SC cells can fire even if there is not an exact spatial and temporal match in connected unisensory areas, providing a mechanism for spatiotemporal tolerances. Evidence from both human and animal research suggests that MSI cells in the SC need exposure to a structured sensory environment in order to develop (Wallace & Stein, 2001; Wallace et al., 2004a; Putzar et al., 2007; Wallace & Stein, 2007; Xu et al. 2012).

1.2.2 Spatiotemporal Congruency in XR is Insufficient

In XR environments, the certainty of time and location estimations for both visual and auditory stimuli may be lower than real environments. The presence of a variety of perceptual issues suggests that spatiotemporal congruency will in many cases be less informative in XR than it is in the real world.

Sound source localization tends to be less accurate when spatial audio is used, compared to localization of sounds that originate from a true point source. Although simulating the binaural disparities of interaural time difference and interaural level differences is achievable by even the most basic spatial audio systems, elevation perception and front-back disambiguation requires simulating spectral changes that occur due to minor differences in the way sounds are occluded by a person's shoulders, head, and ears depending on which direction they come from. These spectral cues can be simulated using Head Related Transfer Functions (HRTFs; Begault, Wenzel, & Anderson, 2001). HRTFs are most effective if customized to fit the shape of each person's outer ear and head/shoulders, but this is typically not feasible. Generalized HRTFs tend not to allow for sufficiently accurate sound source localization, in particular along the azimuth (Wenzel, Arruda, Kistler, & Wightman, 1993). Even with individualized HRTFs created using in-ear microphone recordings, participants exhibit twice as much localization error compared to real-world sounds (Bronkhorst, 1995).

Similarly, although coarse simulation of sound intensity falloff with distance, direct-to-reverberant energy ratio, near-field effects, and spectral changes with distance (Zahorik, 2002) are may be feasible for XR systems, accurate simulation of sound

propagation (e.g., Savioja, Huopaniemi, Lokki, & Väänänen, 1999) is often not. This is in part due to limitations in the computing power available to wearable devices, and in part due to the difficulty of accurately mapping local space in order to simulate reverberations and reflections (Raghuvanshi, Narain, & Lin, 2009).

There may also be delays between changes in the orientation of the HMD and updates to the aforementioned spatial audio effects. Sound playback latency in general presents another impediment to the usefulness of spatiotemporal congruency in XR (Brungart, Simpson, & Kordik, 2005).

A related and less-studied perceptual problem is one of perceived sound source distance and externality/internality. Sounds played over headphones are often perceived as intracranial (Blauert, 1997). Even if direction can be successfully conveyed, conveying the perception of externality is a difficult problem (Iwaki & Chigira, 2016). Sounds perceived as internal may not be perceived as spatially congruent with visual depictions of XR objects located some distance from the user's head.

There are parallel issues with visual perception of the distance of virtual objects, which tends to be underestimated or otherwise ambiguous (Knapp & Loomis, 2004), due to problems such as vergence-accommodation conflict and the difficulty of rendering object reflections at the proper depth. Other visual perceptual issues, such as inaccurate color perception, inaccurate object ordering, latency, tracking error, and issues with object configural goodness stemming from resolution limitations, can lead to visual localization uncertainty (Kruijff, Swan, & Feiner, 2010).

Due in part to these limitations of XR visual and auditory stimuli, the crossmodal binding that occurs with virtual objects tends to have larger spatial binding windows

compared to real-world objects, which could lead to errors in applied multi-object settings. Kytö, Kusumoto, and Oittinen (2015) found that, for virtual objects, the spatial binding window was 5-15 degrees larger than it typically is for real sound sources. The authors did not observe a near-100% rate of "different" responses until 60 degrees of disparity. They recommended at least 30 degrees of disparity be maintained in VEs in order to avoid erroneous binding. Honbolygó, Veller, & Csépe (2012) observed spatial ventriloquism in VR with 10 degrees of spatial separation. Berger et al. (2018) also found that spatial binding windows were much larger in VEs.

Thus, spatial congruency is unlikely to be as useful for XR objects as it is for real objects. Other types of crossmodal congruency, that leverage congruency between the *content* of sounds and visuals, should be leveraged to compensate. Current XR systems do this to a limited extent through the use of pre-selected sound samples, selected by a designer, in the manner of desktop user interface design. However, even well-designed sound samples cannot always be congruent with the variety of possible visual events that may need to be presented to a user in XR. In order to suggest ways in which this situation could be improved, the remainder of this document: (1) identifies a gap in our current understanding of crossmodal congruency, (2) proposes a new set of congruency effects that may be leveraged to improve MSI in XR, and (3) demonstrates the existence of these effects through a human-subjects experiment.

1.3 Associative Congruency

Shams and Kim (2010, p. 296) wrote that "consistency in time, space, structure… and semantics" all contributed to the binding process. *Associative congruency* is here

defined as the set of congruency types to which adherence is determined by prior expectations and beliefs about the world and its contents. "Crossmodal correspondences" is a related and commonly used term that refers more broadly to features in different modalities that seem to correspond (Calvert, Spence, & Stein, 2004; Spence, 2011). Associative congruency is used here to refer to the subset of audiovisual crossmodal correspondences that are: (a) of consequence to MSI on a perceptual level rather than solely on a decisional level; and (b) acquired through experience with perceptual environments. The following sections substantiate these two key aspects of the definition of associative congruency, both of which are built upon relatively recent research in the field.

1.3.1 *Associative Congruency Effects are Direct Perceptual Effects*

Associative congruency affects MSI directly on a perceptual (rather than decisional) level, in a manner set apart from effects on attentional orienting. In a landmark series of studies, Parise and Spence (2008; 2009) used TOJ tasks to find that frequency: size associative congruency could modulate the magnitude of temporal ventriloquism. McGovern, Roudaia, Newell, and Roach (2016) found that adherence to associative congruency could narrow or broaden spatiotemporal binding windows, which the authors suggested could be due to direct modification of low-level tolerances (e.g., Burr et al., 2009), and/or to summation in a parallel accumulator (see Stevenson, Wallace, and Altieri, 2014).

Associative congruency tends to be most impactful at moderate spatiotemporal disparities (Parise & Spence, 2009). If crossmodal stimuli are highly spatiotemporally congruent or incongruent, associative congruency may not have a detectable effect (Keetels

& Vroomen, 2011). Facilitative effects have been observed with stimuli that were 300 ms apart in time, indicating a window of at least 300 ms in which associative congruency effects may have an impact for spatially congruent stimuli, during which crossmodal objects are formed and updated (Chen & Spence, 2017).

A certain level of configural goodness may also be necessary for associative congruency effects to be observed. Martino and Marks (1999) found that frequency: luminance congruency effects ceased to be detected when the figural goodness of visual stimuli was reduced through the utilization of ill-formed, broken-up visual objects.

Notably, although evidence suggests that associative congruency effects are in large part direct perceptual effects, it should be noted that associative congruency also functions indirectly by orienting attention toward congruent audiovisual objects (Koelewijn, Bronkhorst, & Theeuwes, 2010; Chiou, Stelter, & Rich, 2013; Macaluso et al., 2016).

1.3.1.1 Neuroscience of Associative Congruency

The direct perceptual effects of associative congruency are realized through a network of brain regions at several levels of processing. Work by Driver and Spence (2000), and Calvert and Thesen (2004) suggested that MSI involved activity in multisensory convergence areas, which would then influence both early lower unisensory areas and conflict resolution areas. Subsequent research has identified distinct associative congruency areas, as well as feed-back connections through which these areas influence basic spatiotemporal tolerances.

Medial temporal areas, in particular the superior temporal sulcus (STS) and superior temporal gyrus (STG), are involved with associative congruency, with activity occurring a few hundred milliseconds post-onset, suggesting perceptual rather than decisional processes (Belardinelli et al., 2004; Froyen, Van Atteveldt, Bonte, & Blomert, 2008; Noppeney et al., 2007; Olson, Gatenby, & Gore, 2002; Barraclough et al., 2005; Evans, 2007; Su, 2014; Doehrmann & Naumer, 2008; Schneider, Debener, Oostenveld, & Engel, 2008).

In addition to those mediotemporal areas areas, intraparietal areas are also involved with associative congruency effects (Rohe & Noppeney, 2015; Spence & Parise, 2012). Disruption to these areas can modify spatiotemporal tolerances (Bien, Ten Oever, Goebel, & Sack, 2012; Zmigrod & Zmigrod, 2015) or disrupt associative congruency effects (Pourtois & de Gelder, 2002).

Finally, inferior frontal areas act to resolve direct semantic conflicts between crossmodal stimuli (Laurienti et al., 2003; Hein et al., 2007), or otherwise disambiguate ambiguous stimuli (Lundström et al., 2018), as well as enacting top-down factors such as explicitly held expectations (Rahnev, Lau, & de Lange, 2011) or the effects of selective attention (Rahnev, 2017b).

Importantly, these MSI systems contain feed-back projections that act to widen spatiotemporal tolerances when associative congruency is maintained (Clavagnier, Falchier, & Kennedy, 2004; van Wassenhove & Schroeder, 2012; Bhat, Miller, Pitt, & Shahin, 2014; Keil & Senkowski, 2018).

1.3.1.2 Manner in which Associative Congruency Effects Interact

Although a variety of associative congruency effects have been observed (see section 1.3.3), the manner in which adherence to multiple types of associative congruency exert combined influence on perception is not comprehensively understood, and needs to be in order for congruency effects to be effectively utilized in XR. Research in this area can be divided into accounts in which different congruency effects are additive in their overall influence on perception, and accounts in which they interact in more complex ways.

Jonas, Spiller, and Hibbard (2017) evaluated three models of associative congruency feature interaction, and found support for a simple additive model. Participants were asked to specify which auditory stimuli "went with" each visual stimulus. Visual stimuli varied in terms of luminance, color saturation, size, and vertical position. Auditory stimuli varied only in terms of frequency. Using these stimuli, the authors assessed three models of feature interaction. The *summation model* (e.g., Shams & Kim, 2010; Trommershauser, Kording, & Landy, 2011) postulates that congruency adherences are combined in the manner of simple addition, with adherence in terms of more, and more heavily weighted, features leading to stronger crossmodal integration. The *hierarchy model*, in line with the conclusions of Melara and Marks (1990, see below), predicts that congruency types have a hierarchy of predictable interaction effects, in which adherence to or violation of some types of congruency will dominate others. Lastly, the experimenters evaluated the *majority model,* which is a simplified linear summation model that posits that the integration decision can be modeled as going to whichever visual object was congruent via the greatest number of features (which are all equally weighted). The authors found that the linear summation model best accounted for the audiovisual pairings indicated by participants.

Research in the area of cue competition and the neural encoding of probability can provide a model for how the cue weighting required by the linear summation model could be realized. Powell, Merrick, Lu, and Holyoak (2016) found that (unisensory) priors compete to influence perception. Participants were trained on a set of trials in which three vegetables were depicted, at least one of which they knew had caused a person to get sick. This created weak priors on those vegetables (cues). Cue A had a true sickness-causing strength of 0.5, cue C 0.9, and cue B either 0.2 or 0.8. Participants were then asked to assess the likelihood that cue A made a person sick. However, cue A had never been presented on its own during training, so its true causality was unknown. Responses indicated that the strength of cue B had affected the prior on Cue A. In two additional experiments, the authors had the cues occur independently, and then had one cue predict the opposite polarity (that is, predict a person not becoming sick). In both cases, similar prior competition was observed. These findings suggest that different associative congruency features may have different reliabilities, and may be weighted differently depending on those reliabilities, as well as the reliabilities of other congruent or incongruent features. Modeling studies such as Ursino, Cuppini, and Magosso (2017), who demonstrated that a Hebbian learning computer model could produce behavior in which multisensory signals were weighted in a fashion that compensated for sensory noise by utilizing acquired priors, provide convergent evidence. Other studies have identified various mechanisms by which populations of neurons could encode cue reliability (Knill & Pouget, 2004; Bach & Dolan, 2012; Hartmann, Lazar, & Triesch, 2014; Pitkow & Angelaki, 2017; Rahnev, 2017a). Activity in mediotemporal and prefrontal MSI areas has also been shown to vary in accordance with the spatiotemporal certainty of stimuli (Nastase, Davis, & Hasson, 2018).

Despite the results of Jonas, Spiller, and Hibbard's 2017 study, there have been documented cases in which interactions have been found between associative congruency effects. Melara and Marks (1990) used a two-dimensional speeded classification method to investigate the way related auditory parameters interacted. The authors found evidence that some dimensions are "hard," or resistant to Garner interference, whereas others are "soft," or vulnerable to Garner interference. Sound frequency was "hard," and loudness "soft." The authors suggested that certain sound attributes, such as frequency, were more fundamental to an object's identity than others. Bonetti and Costa (2018) discovered a dimensional interaction between frequency: elevation congruency and frequency: size congruency. Participants were asked to classify the location of a sound as low or high. Sounds were emitted either above or below the level of the participant's ears, and had frequencies of either 100, 200, 600, or 800 Hz. Participants were asked to classify whether the sounds came from above or below. When higher located sounds were also higher frequency, classification was speeded. However, matching low located objects to lower frequency sounds did not facilitate faster reactions. In a second experiment, participants matched tones with circles of various sizes, revealing an inverse relationship between the logarithm of frequency and circle diameter. The authors asserted that a novel interaction was occurring: lower frequency sounds supported the hypothesis that an object was larger, which then decreased the certainty of the sound localization (because the sound could emanate from any part of the object). This decrease in auditory localization certainty was enough to slow reactions, despite low frequency: low elevation congruency being maintained. Pisanski et al. (2017) also found frequency: elevation and frequency: size interactions. Participants were asked to judge the location and size of a higher or lower

frequency voice coming from either high or low locations. Lower frequency voices were rated as being larger, regardless of spatial location. This suggests that frequency: size effects may be stronger than frequency: elevation effects.

Overall, these studies indicate that the linear summation model (e.g., Jonas, Spiller, & Hibbard, 2017), with features weighted by their reliability (e.g., Powell, Merrick, Lu, & Holyoak, 2016) can accurately describe some feature combinations. However, there are cases in which features may interact (e.g., Melara & Marks, 1990), in ways not yet systematically understood.

1.3.2 Associative Congruency Effects Stem from Perceived Regularities

A potential source of feature reliability estimates that could be used to form a weighted sum is the fact that auditory and visual features encountered during daily life are predictive of each-other to varying extents. Rather than being arbitrary, associative congruency effects come about due to implicit recognition of regular co-occurrences present in perceptual scenes.

The phenomenon of perceptual learning of perceptual scene statistics been best documented in the area of unisensory visual perception. Girshick, Landy, and Simoncelli (2011) found a bias toward perception of line orientations as being cardinal, which reflected the distribution of line directions in a sample of natural scenes. In a similar study, Peters, Balzer, and Shams (2015) found that a "smaller is denser" prior was ecologically justifiable, and that that untrained participants had response biases reflecting the ecological correspondence between those two visual features. Gerhard, Wichmann, and Bethge (2013) observed that participants could even implicitly learn statistical regularities

contained within low-resolution, random-seeming texture scrambles. After viewing a set of such images, participants completed a two-alternative forced-choice task and were able to discriminate the training images from the real-world images.

The results of Parise, Knorre, and Ernst (2014) suggest that this type of learning takes place for crossmodal regularities as well as unimodal regularities. The authors recorded a set of natural sounds, as well as the HRTF for each sound. Using a model-fitting procedure, the authors found a statistically significant relationship between object elevation and sound frequency, after removing the impact of HRTFs. Across natural scenes, auditory frequency was a meaningful predictor for object elevation relative to the listener. Thus, frequency: elevation congruency would have utility to a human observer. Munoz and Blumstein (2012) highlighted the potential usefulness of utilizing such crossmodal learning to reduce uncertainty in attempts to detect, localize, and quickly acquire information about objects, especially in degraded perceptual environments such as those that are noisy or dark.

These results concur with ongoing research in developmental psychology, that has found that children are not born with "neonatal synesthesia" (Deroy & Spence, 2013), are less susceptible to crossmodal illusions that leverage associative congruency (Mildner & Dobrić, 2015; Nava, Grassi, & Turati, 2016), and yet are capable of acquiring associative congruency effects through brief training (Thomas, Nardini, & Mareschal, 2017). Additionally, associative congruency effects that would depend upon visual input to be acquire tend to be absent in persons with blindness (Hamilton-Fletcher et al., 2018). These studies suggest that the acquisition of multisensory perceptual experiences facilitates

implicit learning of crossmodal regularities, which in turn gives rise to observable associative congruency effects.

1.3.2.1 Ongoing Perceptual Learning Supports Associative Congruency Effects

Since associative congruency effects are the results of sensory experiences, they should share properties with more widely studied cases of perceptual learning. Perceptual learning can occur with relative ease and rapidity (Gilbert, Sigman, & Crist, 2001), suggesting that humans may be suspectable to far more associative congruency effects than have been documented thus far.

Aslin and Newport (2008) defined perceptual learning as a process through which learners quickly acquire information from the environment without explicit feedback. The brain must be plastic enough to adapt to new perceptual environments, but not so plastic that valuable existing learning is overwritten. This has led to the suggestion that that a given perceptual decision must be related to a person's current task for a reinforcement signal to be sent and learning to occur (Seitz & Watanabe, 2005). Although such task-relevancy may *facilitate* multisensory learning, it has been shown to not be required (Seitz & Leclercq, 2012). Kim, Seitz, Feenstra, and Shams (2009) addressed two other arguments against the pervasiveness of perceptual learning: (1) that rapid perceptual learning may not persist longer than a typical laboratory session and (2) that perceptual learning is explicit learning rather than implicit learning. The authors found that perceptual learning could persist after a twenty-four-hour retention interval, and was independent of explicitly acquired knowledge.

Findings on perceptual learning of unisensory visual priors have indicated that such priors are in fact constantly being acquired, in a manner both rapid and flexible, in response to changing perceptual ecologies. Sotiropoulos, Seitz, and Series (2011) found that a "slow speed" prior, in which perception of movement speed is biased toward slower interpretations, could be modified with a small number of training sessions. In a related study, Chalk, Seitz, and Series (2010) studied the speed with which perceptual learning could take place. Participants were tasked with determining the angle of movement of an array of dots. Unbeknownst to participants, the statistics of the dot movement directions were not uniform: -32°/32° movement directions were six times more likely than other directions. In less than two hundred trials, participants began to bias their reported movement directions toward -32°/32°, and even started to report illusory motion in those directions for trials in which there had been no movement.

It can be concluded from these studies that perceptual learning is highly plastic, allowing human observers to continually integrate a variety of information about natural scene regularities into the manner in which their perception is biased.

Crossmodal perceptual learning appears to be occur in a similarly rapid and ongoing manner. Research has shown that multisensory perceptual learning occurs in large part distinct from, and in parallel to, unisensory perceptual learning relating to the same stimuli (Mitchel, Christiansen, & Weiss, 2014; Paraskevopoulos, Kuchenbuch, Herholz, and Pantev, 2012; Seitz, Kim, van Wassenhove, & Shams, 2007). However, there is some overlap (Mitchel & Weiss, 2011; Glicksohn & Cohen, 2013). Heron et al. (2012) found that participants could acquire a novel sound frequency: visual spatial frequency prior, and experienced corresponding congruency effects. Adherence to this recently-learned type of

associative congruency led to an increase in spatiotemporal tolerances, indicating that MSI was being affected. Habets, Bruns, and Röder (2017) found that simultaneity judgment rates were higher when participants experienced auditory and visual stimuli that had previously occurred more often together during a training phase, compared to other stimuli. Ernst (2007) trained participants on an arbitrary visual luminance: tactile stiffness congruency, and found that participants were subsequently able to discriminate smaller differences in luminance when stimuli were presented with congruent tactile information.

Fifer, Barutchu, Shivdasani, and Crewther (2013) compared perceptual learning for audiovisual pairings versus visual-visual pairings. The authors found that learning progressed more quickly for the multisensory pairings, and attributed these results to the facilitative properties of bound multisensory stimuli. The authors contrasted this result with the findings of Tanabe, Honda, and Sadato (2005) in which visual and auditory stimuli were presented with a sixteen-second offset, suggesting that crossmodal perceptual learning occurs most easily when temporal congruency is at least moderate (e.g., Spence & Squire, 2003).

Remarkably, several studies have found that, in controlled conditions, subsequent crossmodal perception of an audiovisual object can be modified with only a single exposure to co-occurrent stimuli (Van Wanrooij, Bremen, & John Van Opstal, 2010; Wozny & Shams, 2011).

Piazza, Denison, and Silver (2018) found that newly acquired associative congruency effects could influence conscious visual perception. First, participants were passively exposed to a series of arbitrary audiovisual pairings, alongside auditory tones.

Next, participants completed a binocular rivalry task, in which a different (equally familiar) image was shown to each eye, and indicated what they saw. When participants heard the tone that had been previously co-occurrent with the image shown to one eye, they were significantly more likely to report seeing that image.

These studies suggest that crossmodal perceptual learning is continually occurring, in the wide variety of cases in which crossmodal co-occurrences are present in the environment. As such, rather than crossmodal congruency effects being reserved for a subset of fundamental "synesthetic" auditory and visual parameters, or stimuli that are explicitly semantically matched, it is instead likely that there are a variety of heretofore unexplored congruency effects. The next section describes two existing types of associative congruency, and suggests a third type to account for many such possible effects.

1.3.3 Types of Associative Congruency

Associative congruency can be divided into three types: *semantic congruency*, *parametric congruency*, and *action-object congruency*. Parise (2012) described semantic congruency as specific learned associations, and parametric (or "synesthetic") congruency as mappings between fundamental auditory and visual/object dimensions. Added to these is a proposed new type called action-object congruency, which refers to crossmodal congruency between the sounds and visual appearance of sound-producing events.

1.3.3.1 Semantic Congruency

Highly specific associative learning that influences multisensory perception can be characterized as leading to *semantic* congruency effects. Semantic congruency can be

defined as congruency between auditory and visual stimuli that have been *recognized*– that is, assessed as a whole, rather than on any specific feature or parameter– and associated with a concept recalled from memory.

A variety of studies have shown that semantic congruency between auditory and visual stimuli facilitates faster processing of those stimuli (Laurienti et al., 2004; Molholm, Ritter, Javitt, & Foxe, 2004; Yuval-Greenberg & Deouell, 2009; Chen & Spence, 2010). Semantic congruency affects crossmodal binding, as evidenced by broadened temporal binding windows (Ten Oever et al., 2013). The majority of neuroscientific research on MSI (e.g., Doehrmann & Naumer, 2008) has used semantic congruency of stimuli (often, animal sounds and visuals) as a means of studying the broader phenomenon of MSI.

Semantic congruency more-so involves late processing compared to other types of associative congruency (Spence, 2011), has a prominent decisional component (Koppen, Alsius & Spence, 2008), and there may be unique effects for certain types of stimuli (Suied & Viaud-Delmon, 2009). Determination of semantic congruency involves the assessment of explicit crossmodal *conflicts*, and has tended to be associated with activity in frontal areas typically involved with conflict resolution, as opposed to lateral temporal areas (Doehrmann & Naumer, 2008; Su, 2014).

Semantic congruency is also distinct from other types of associative congruency due to the inclusion of congruency between complex constructs, such as apparent visual gender and apparent auditory gender. Smith, Grabowecky, and Suzuki (2007) had participants classify androgynous faces as either male or female. When faces were paired with pure tones in the typical male fundamental frequency range, the faces were more likely

to be classified as male; the inverse was true as well. Vatakis and Spence (2007) also showed that binding was more likely to occur when a complex, multi-feature semantic congruency type was maintained. Participants saw visual mouth movements and heard spoken syllables with a slight temporal asynchrony, and completed a TOJ task. JNDs were smaller when the auditory and visual gender of the speaker were incongruent. In a similar study, Van Wassenhove, Grant, & Poeppel (2007) found that congruent or incongruent McGurk stimuli led to a difference in simultaneity judgment rate.

Also distinguishing semantic congruency are the fact that: (a) different types of semantic constructs lead to different effects on MSI; and/or (b) feature complexity moderates the magnitude of semantic congruency effects.

Suied and Viaud-Delmon (2009) found evidence for the former possibility. The authors compared response times for classification of images of either animals or types of transportation, in the presence of irrelevant distractor sounds that could be either animal or transportation sounds. When auditory and visual stimuli were congruent, content type did not have an effect on response times. However, when stimuli were incongruent, and the auditory stimulus was an animal sound, response times were higher compared to when the stimulus was a transportation-related sound. The authors suggested that irrelevant animal sounds were more difficult for prefrontal areas to inhibit due to the historic importance of orienting to animal threats.

Other studies have suggested that the complexity of evoked semantic constructs is a more likely explanation for differences in the impact of semantic congruency, precluding the need for content-effect-based explanations. A body of research into the "Colavita visual

dominance effect," which occurs when participants tend to respond to the visual aspect of multimodal targets rather than the auditory component, speaks to this possibility. Koppen, Alsius, and Spence (2008) utilized a speeded classification procedure (participants were asked to indicate which modality contained a cat/dog) and found that semantic congruency had no impact on the Colavita effect. The authors noted that prior studies (in which the effect had been found) had used simple artificial stimuli such as pure tones, as opposed to the more complex speech and animal sounds the authors had used. In a second experiment, Koppen, Alsius, and Spence (2008) instead asked participants to identify the multisensory object, and the Colavita effect re-emerged. The authors asserted that more complex cases of semantic congruency determination more-so involve later processes, and thus do not influence perceptual phenomenon such as the Colavita effect.

Yuval-Greenberg and Deouell (2009) suggested that MSI systems weight more complex features, as well as features perceived with greater certainty, more heavily in determining the impact of semantic congruency on the ultimate percept. They had participants identify visual and auditory stimuli (animal sounds and visuals), and found that the speeded classification effects of adhering to semantic congruency were smaller when the visual stimuli were lower in contrast, and thus more difficult to perceive.

Speech has been indicated as either a unique type of semantic congruency, or as an example of a complex, configural feature that is weighted more heavily than simpler, less informative features. Vatakis, Ghazanfar, and Spence (2008) had participants complete a TOJ task for both human speech and rhesus monkey sounds that were experienced alongside visual stimuli that were either congruent or incongruent. The authors found that temporal ventriloquism effects were present for human speech sounds, but not for rhesus

monkey sounds, and suggested that human speech was unique in terms of congruency effects. Vatakis and Spence (2006) found that the use of more complex speech stimuli led to increased crossmodal binding and wider temporal tolerances compared to less complex stimuli, via a TOJ task. They attributed this to there being more spatiotemporal "sub-events" in the speech stimuli, and/ to more complex stimuli leading to stronger semantic congruency effects overall.

These studies characterize semantic congruency as comparisons between recognized visual objects and sounds, and as comparisons between complex, multifaceted constructs. In both cases, explicit comprehension and recognition of stimuli appears to be required.

1.3.3.2 Parametric Congruency

Parametric congruency refers to broadly-applied congruency effects between low-level auditory and visual *parameters*, such as auditory frequency and visual size. Such crossmodal correspondences have been characterized as "synesthetic congruency," referring to the theory that parametric congruency shares mechanics with synesthesia (Spence, 2011). Persons with synesthesia do tend to experience parametric congruency effects (Sagiv & Ward, 2006; Ward, Huckstep, & Tsakanikos, 2006), but it is now known that parametric congruency effects are different from the crossmodal effects experienced by persons with synesthesia (Parise & Spence, 2013). Persons with synesthesia also tend to experience a variety of crossmodal congruency effects not found in non-synesthetes (Chiou, Stelter & Rich, 2013). Like other types of associative congruency, parametric congruency effects instead come from sensory experience, and reflect recognized

regularities of a person's perceptual environment (Parise, Knorre, & Ernst, 2014). Parametric congruency affects MSI directly, on a perceptual level (Parise, 2012).

A variety of parametric congruency effects have been found. One of the most widely researched effects is the relationship between auditory frequency and visual elevation, with more highly elevated visual stimuli being congruent with higher frequency sounds (Pratt, 1930, Mudd, 1963; Roffler & Butler, 1968; Bernstein & Edelstein, 1971; Melara & O'Brien, 1987; Ben-Artzi & Marks, 1995; Evans & Treisman, 2010; Bonetti & Costa, 2018). This effect depends on a person's perceptual upright (Carnevale, 2015; Carnevale & Harris, 2016), and does not require musical training (Rusconi et al., 2006).

Parametric congruency effects between visual size (specifically, subtended visual angle) and auditory frequency have also been observed consistently (Sapir, 1929; Evans & Treisman, 2010; Gallace & Spence, 2006; Boyle, Kayser, & Ince, 2018). Parise and Spence (2008) found direct evidence that frequency: size parametric congruency could influence the magnitude of temporal ventriloquism and thus affect audiovisual integration. Adhering to frequency: size congruency can also assist with motor planning (Rinaldi et al., 2016), and can improve timing perception for children with dyslexia (Chen et al., 2016).

Parametric congruency effects between frequency and luminance have also been observed (Marks, 1974; Hubbard, 1996; Ludwig, Adachi, & Matsuzawa, 2011). Some of those congruency effects may be due in part to structural similarity between the neural encoding of luminance and frequency (Spence, 2011; Chan & Dyson, 2015). Other parametric congruency effects include timbre and chroma (Ward, Huckstep, & Tsakanikos 2006; Hamilton-Fletcher et al., 2018), the "roundness" of phonemes and visual

curvilinearity (Westbury, 2005; Makovac & Gerbino, 2010; Parise & Spence, 2012), waveform curvilinearity and visual curvilinearity (Parise & Spence, 2009; Parise & Spence, 2012), and auditory tempo and visual spatial frequency (Guzman-Martinez et al., 2012).

Whereas semantic congruency operates on well-formed, complex objects associated with specific expectations, parametric congruency operates on the level of fundamental visual and auditory features (i.e., parameters). Unlike holistic semantic conflicts (e.g., Chen et al., 2018), participants may not be able to articulate conflicts between parameters, even as they show behavioral evidence of parametric congruency effects. Melara (1989) suggested that parametric congruency effects occur through a separate intersensory stimulus formation process, distinct from processing of semantic congruency.

Parametric congruency effects are also characterized by their ability to be readily described in terms of an underlying dimensional space. This can be contrasted with semantic and action-object congruency, which leverage perceptual expectations that are tied more closely to the specific stimuli in question. Karwoski, Odbert, & Osgood (1942) proposed that crossmodal correspondences could be the result of repeated co-activation between basic features in different sensory modalities, that could be described as reflecting an underlying parameter space. Walker, Walker and Francis (2012) conducted an experiment in which participants were asked to rate various visual, auditory and tactile stimuli on different semantic differential rating scales. The authors found evidence of an underlying dimensional space inhabited by corresponding stimulus dimensions in separate modalities.

Another distinguishing feature of parametric congruency effects is that they may be relative rather than absolute. Relative effects change depending on which parameter values have been experienced recently, whereas absolute congruency effects are consistent if a presentation is experienced in isolation. Marks (1974) found that higher frequency tones played alongside lighter stimuli, or lower frequency tones played alongside played alongside darker stimuli, led to speeded classification compared to when congruency was violated. This suggests that frequency: luminance congruency effects are relative rather than absolute. In a series of experiments, Walker and Walker (2016) found that a visual brightness: tactile size congruency effect also performed in a relative fashion, with the same brightness corresponding to different tactile sizes in different experiments, depending on the surrounding context. In a similar study, Brunetti et al. (2018) had participants classify visual stimuli as large or small as they heard lower, medium, or higher frequency task-irrelevant sounds. The "medium" frequency sound would have been "higher' if it followed the low frequency sound, or "lower" if it followed the high frequency sound. The authors found that the medium-frequency sound changed its effect depending on which sound had preceded it, again indicating that some parametric congruency effects are better characterized as being relative rather than absolute.

1.3.3.3 Action-Object Congruency

Rather than considering the general co-occurrence of fundamental auditory and visual parameters, humans may also be able to determine crossmodal congruency in terms of mid-level auditory and visual features specific to individual sound-producing events. Research on the ability of humans to extract varied information about crossmodal objects, through the sounds produced by physical events, suggests that humans can indeed extract

mid-levels features of this nature, and aforementioned evidence has shown that perceptual learning can occur with only a few exposures (e.g., Wozny & Shams, 2011). Humans with typical auditory and visual perceptual abilities, raised in typical sensory environments, will have had a vast number of experiences with physical objects and sound-producing events. Perceptual learning is likely to have occurred based on the auditory and visual features that could be extracted from those events. Thus, congruency effects should exist that derive from the expected auditory and visual behavior of sound-producing physical events. These are here referred to as *action-object congruency* effects.

A key difference between action-object congruency and semantic congruency is that the former is conceived as the result of comparisons between mid-level auditory and visual feature estimates, which themselves stem from properties of sound-producing events, whereas semantic congruency refers to holistic evoked constructs. An entirely novel virtual object could adhere to action-object congruency, based on its apparent properties, even if the listener did not have experience with the sounds that object produces. This distinction is especially important to make in when considering virtual objects, that may be utterly unfamiliar and thus semantically uncertain.

As such, this category of effects could be particularly impactful for XR. The findings of Bailey, Mullaney, Gibney, and Kwakye (2018) that more realistic virtual objects were more likely to be bound suggests that defining and adhering to action-object congruency could assist with solving the correspondence problem in XR. However, current XR systems tend not to leverage the subtleties of matching sounds to the particulars of simulated sound-producing events. Suggested methods for creating adherent sounds and visuals are discussed section 1.3.3.9.

Another important difference between action-object congruency and semantic congruency is that action-object congruency relates to the manner which a perceived sound-producing *event* occurs in each modality. By contrast, semantic congruency leverages evoked associations relating to discrete category membership, which can occur without a perceptible sound-producing event. For example, semantic congruency effects may emerge when an image of an animal is displayed alongside the sound made by that animal, but the *production* of the sound need not be displayed. Action-object congruency requires that the perceiver see and hear a sound-producing event to extract the necessary features. Determining action-object congruency is conceived as a process of *rapid evaluation* of event features rather than holistic *recognition* of auditory and visual events, in the manner of semantic congruency, or the broad application of general associations to any situation, in the manner of parametric congruency. Making a distinction of this sort was suggested by Connolly (2014), who advocated a shift from discussions of crossmodal binding of stimuli based primarily on semantic congruency, to a "unitization" view focused on sound-producing events, which the author characterized as a process distinct from semantic accounts relating to category membership, and from low-level parameter-based accounts.

If parametric congruency can be characterized as the application of highly generalized perceptual learning relying on low-level parameter comparison, and semantic congruency as highly specific statistical learning relying on complex, holistic objects, then action-object congruency can be described as stemming from mid-level feature comparisons related to the nature of sound-producing events themselves. Considering action-object congruency means considering the ability of perceptual processing to rapidly

perceive mid-level features of crossmodal events, in both the auditory and visual modalities, in order to facilitate determination of the congruency of those features based on past perceptual experience.

Action-object congruency can be divided into *action congruency* and *object congruency*. This reflects a demarcation similar to the that suggested by Gaver (1993), who divided the physical properties derivable from sounds into *object* and *interaction* properties, as well as the action/object paradigm explored by Conan et al. (2013).

1.3.3.4 Object Congruency

Gaver (1993) described how sounds encode a variety of pieces of information about the objects involved in sound-producing events. Key derivable features include the material of the objects, resonator size, and surface rigidity. Steenson, Rodger, and Matthew (2015) suggested that sounds should be conceptualized as evidence of material interactions rather than solely as "sounds." Preis and Klawiter (2005) proposed a three-level account of auditory information processing, with derivation of object-related information as a key component. Their three levels were: (1) auditory stream segregation, (2) localization of a sound sources/ spatial processing, and (3) determination of the properties of sound-producing objects. The following sections detail object parameters that evidence suggests humans can derive from the sounds produced by physical events.

1.3.3.5 Object Congruency Features

1.3.3.5.1 Object Size

A basic derivable object feature is object size. Von Kriegstein and Giraud (2006) noted that the frequency content of emitted sounds depends on the size and type of an object's resonator, and showed that humans were sensitive to these differences. Participants in their study were able to identify the size of humans, French horns, and bullfrogs, based on their different frequency profiles. Notably, these were complex, ecological sounds rather than simple higher or lower frequency tones. Neuroimaging revealed activation in the STG as well as anterior temporal areas. The authors suggested that the STG stores information relating to acoustic scale for human speech, and that areas within the anterior temporal lobe process general acoustic scale information and determine audiovisual congruency for those features.

Grassi (2005) had participants listen to the sounds of different-sized balls being dropped onto different-sized plates, and then estimate the size of the balls. Participants were able to scale their reproductions of the size of the dropped ball with reasonable accuracy. However, when balls were dropped onto larger plates, participants tended to overestimate the size of the ball, since the plate was producing most of the sound. Grassi, Pastore, and Lemaitre (2013) repeated this procedure, but introduced inaccurate amplitude or frequency components. Participants remained able to judge ball size based on the produced striking sound, indicating that both frequency and amplitude are utilized to determine the size of objects based on sound.

Lakatos, McAdams, and Caussé (1997) found that listeners were capable of telling the difference between different-sized bars that were struck, utilizing the fact that differing torsion and bending of the bars led produced different sounds. Coward and Stevens (2004) found a difference between learned and pre-existing action-object expectations using the same domain of study. The authors had participants listen to impact sounds (pipes being struck) and identify the size of the pipes. Participants were better at making this estimation when expected "nomic" mappings were used (frequency: pipe length) compared to when novel "symbolic" mappings (damping: pipe length) were trained in the study.

Listeners can also determine object density and composition, in addition to pure size. Lutfi (2001) demonstrated how a listener could theoretically tell whether a struck object was solid or hollow, and found that participants were able to do this with moderate accuracy. Pisanski et al. (2017) found that a frequency: resonator size mapping was present for ecological sounds (human voices) that differed via subtle manipulations to frequency content, as well as for ecological judgments of "body size."

Human listeners also tend to be able to extract information about the size of rolling or bouncing objects based on sound. Stoelinga (2007) conducted a series of experimental and model-building activities that explored the sounds made by rolling and bouncing objects. The author referred to the study of physics in conjunction with perceptual information as "psychomechanics," and identified two such pieces of information that are encoded in the sound of a rolling object. First, the "restitution coefficient," defined as the proportion of time intervals between object bounces, could be used to determine the size of a bouncing and rolling ball. Second, the spectral content of impacts made by the ball can be informative of the same property. The author found that spectral content was more

impactful on size determination than the restitution coefficient, supporting the principle that amplitude patterns generally convey information about actions, and that frequency generally contains information about objects. In a similar study, Cabe, Bochtler, and Neuhoff (2018) found that participants were able to determine the size of rotating elliptical objects based on the sounds they made, again by utilizing spectral cues. Rotating ellipses make predictably variable sounds as different parts of the ellipse edge come into contact with a nearby flat plane. The authors found that untrained participants were able to correctly discriminate between larger and smaller ellipses based on this sound property.

The aforementioned studies suggest a pervasive human ability to recognize the size of objects via the sounds produced by those objects. This information should thus be represented in acquired multisensory priors, and congruency effects should exist.

1.3.3.6 Material Type

Another component of object congruency is congruency of perceived material type, which can be determined via cues such as amplitude envelope and timbre. Amplitude envelope (the timing and magnitude of attack, decay, sustain, and release phases) varies based on the physical properties of the materials involved in a sound producing event, and can be used to distinguish material type. Cazabon (2016) found that the amplitude envelope of sounds influenced the extent to which spatial ventriloquism occurred between those sounds and visual stimuli. The author established a discrimination threshold for amplitude envelopes, then played a sound either above or below that threshold, alongside a differently-located visual flash, and tasked participants with localizing the sound. For sounds with amplitude envelopes in the higher group, participants did not exhibit signs of

spatial ventriloquism; however, sounds with lower amplitude envelopes (less than 16 ms) did lead to spatial ventriloquism.

Chuen and Schutz (2016) had participants perform a TOJ task in which the sounds used either had natural high or low amplitude envelopes. In one case a naturalistic recording of a bowed instrument (cello) was used, and in another case a recording of a struck instrument (marimba) was used. Visual stimuli were either matched (same instrument) or unmatched (different instrument). When auditory and visual stimuli were matched, TOJ performance was worse than when they were not matched, indicating stronger crossmodal binding. Subsequently, the authors removed spectral cues (timbre) for the two instruments. This led to congruency effects becoming nonsignificant. In a third experiment, the authors removed the amplitude envelope differences but preserved spectral differences. This led to significant congruency effects. The authors concluded that amplitude envelope and spectral content are both utilized to determine crossmodal congruency between visuals and sounds, with spectral content being the stronger cue.

1.3.3.7 *Action Congruency*

Gaver (1993) also described an extensive array of information about the nature of sound-producing *actions* that is encoded in the sounds produced by physical events. Subsequent research has isolated a set of such features experimentally and through modeling activities. Conan et al. (2014) found that amplitude variations over time carry information about the type of action that produces sounds, and that humans were able to utilize those features to classify actions. Importantly, this was true even when abstracted, synthetic sounds were used, that did not resemble ecological sounds. The abstracted

elements typical of specific actions are known as *transformational invariants,* and provide direct evidence as to the type of action that produced a given sound (rolling, scraping, breaking, etc.). Although humans have the earlier-described abilities to perceive object properties during such events, Lemaitre and Heller (2012) found that action perception was more robust. This suggests that action congruency effects may be more impactful on MSI compared to object congruency effects.

Auditory perception of bouncing, rolling and breaking actions has been the topic of several studies. Warren and Verbrugge (1984) found that participants could distinguish between bouncing or breaking actions on the basis of amplitude variation patterns derived from the pattern of impacts. Carello, Wagman, and Turvey (2005) conducted an acoustic analysis of the manner in which amplitude patterns reflect action properties, again focusing on distinguishing breaking from bouncing objects. The authors isolated defining amplitude patterns for those actions, and their periodicity. A bouncing object produces a single, damped, quasi-periodic amplitude pattern, whereas a breaking object produces an initial large amplitude spike followed by subsequent damped, quasi-periodic spikes. The authors also described a generalizable method a listener could use to predict how actions would sound, providing a model for how human listeners could classify a variety of different actions based on the amplitude profiles symptomatic of different types of impact sequences. Grassi and Casco (2010) utilized a methodology in which two visual disks moved toward each other, and then could be perceived as either passing through or bouncing off of each other. If a sound was played at the moment of contact, participants were more likely to report perceiving a bounce. If this sound was congruent, participants were even more likely to perceive a bounce than if the sound was incongruent. This study

is an example of action-object congruency being associated with a qualitative change in perception. Parise and Ernst (2017) utilized the same paradigm, and found that "bounce" perception became more likely if auditory and visual stimuli during the object interaction matched expectations about the motion energy that would be present after the bounce. In addition to bouncing and breaking objects, humans are also able to discriminate the speed of rolling objects via sound. Houben, Kohlrausch, and Hermes (2004) asked participants to identify the properties of rolling balls based on their sounds. Participants were able to discriminate between high and low speed rolling balls. These studies indicate that the expectations behind action congruency are more granular than the *categorically* correct action being represented through visuals and audio. Instead, matching the details of each sound-producing action is likely to impact MSI.

Other research has revealed an ability to distinguish between rubbing and scraping actions. Conan et al. (2012) found that participants could distinguish between various recordings of rubbing or scraping actions. Building on that work, Conan et al. (2013) isolated the transformational invariants that delineated those two action types. Rubbing actions are typified by a high-density series of impacts between the rubbing object and various surface irregularities of the rubbed object. The net outcome of those many small impacts is a sound with relatively constant amplitude. By contrast, scraping actions produce sporadic, lower-density impact patterns. This tends to produce waveforms characterized by high peaks and low troughs, as well as variability in peak timing. Thus, impact series/ amplitude modulation patterns can be used to distinguish rubbing and scraping. The authors developed a quantitative model that related action properties (rub-ness, scrape-ness, or roll-ness), and then developed a synthesizer that could simulate those

action-specific amplitude patterns, in addition to object properties such as material and shape.

Action congruency effects have also been found in the domain of musical instrument sounds. Graham (2017) assessed the existence of a "musical McGurk" effect. Vibrato comes from the action of rapidly moving parts of musical instruments. Musicians are likely to have specific associations about the amount of auditory vibrato that should be produced by the instrument, for a given visually perceived amount and speed of that movement ("visual vibrato"). The authors showed video of cellos being plucked/bowed, as well as trombones being played, to participants with musical experience. Simultaneously, they played an auditory note with varying degrees of congruency in terms of the amount of vibrato. The authors found evidence of action-object congruency effects (modified perception of vibrato amount) for these stimuli.

It should be noted that action congruency is more than just congruency effects between auditory temporal structure and visual temporal structure (Parise, Spence & Ernst, 2012; Parise, Harrar, Ernst & Spence, 2013; Nidiffer, Diederich, Ramachandran, & Wallace, 2018). Su (2018) observed that congruency of temporal structure can affect integration for more complex stimuli as well (abstract figures dancing the Charleston). Although this type of fundamental temporal structure congruency is likely a prerequisite, action congruency also involves congruency between specific details of the sound-producing event, in addition to their basic temporal structure.

One such detail is the timbre differences associated with different scraping actions. Thoret et al. (2014) found that participants were able to reconstruct drawn shapes based on

the friction sounds produced by the movement of a pencil across paper. The investigators found that sound timbre varied with velocity profile. Drawing a circle produces a constant velocity and timbre, whereas drawing an ellipse produces a sinusoidal velocity profile and according timbre pattern. Participants were able to discriminate between shapes after hearing synthetic sounds that contained those timbre patterns but did not directly resemble the sound of a pencil upon paper.

1.3.3.8 Feasibility of Visual Feature Extraction

The aforementioned studies indicate that humans can determine a variety of action and object features via the sounds that are produced. Visually, many of those same features can be determined via brief assessment. In addition to fundamental abilities to perceive the motion of interacting objects, humans can readily discern object features such as shape (Hummel, 2000) and size relative to their perceived body (Van Der Hoort & Ehrsson, 2016). Humans are also able to visually estimate object features such as texture (Lee & Sato, 2001; Tiest & Kappers, 2007), and surface softness/stiffness (Wu, Basdogan, & Srinivasan, 1999; Cellini et al., 2013). As such, determination of action-object congruency should be feasible through a process of brief auditory and visual assessment, without explicit recognition of interacting visual objects or the produced sound.

1.3.3.9 Action-Object Sound Design and Simulation

A practical approach to creating action-object adherent XR scenes would be to provide a larger palate of pre-recorded sounds, and then select from them dynamically depending on the specifics of each visually depicted sound-producing event. In parallel, the depiction of instances of sound production in XR as physically-caused sound-producing

events, as opposed to abstract or event-less sound emissions, could take advantage of action-object congruency effects.

In addition to those measures, XR designers could leverage a number of systems have been developed that synthesize sounds from scratch based on simulations of sound production. Integrating such systems into XR systems could facilitate "automatic" adherence of generated sounds to action-object congruency, even if virtual objects, and the actions undertaken with them, do not resemble anything that exists in the real world. Some models simulate the full physics of sound production, whereas others are designed to produce the type of action-object features that can be recognized by human listeners, without requiring high-fidelity sound production simulation.

Darvishi et al. (1995) laid out a sound synthesis framework in which sounds would be generated based on the physical properties of the objects involved, which entailed synthesizing the initial waveform itself based on granular models of object shape and surface geometry, as well as the type of action. Van Den Doel, Kry, and Pai (2001) as well as Rocchesso (2004), and Mullan (2009) used similarly detailed collision physics to generate sounds.

Stoelinga and Lutfi (2011) took a different approach, and proposed a parsimonious model of object impact, bouncing and rolling sounds, that was based on perceptual needs rather than exhaustive physical simulation. This model produced auditory features of the sort that Stoelinga (2007) had showed were useful in making determinations about the nature of object motion, without delving into full physical simulation of sound production. Produced sounds were optimized to contain information useful to human perception of

bouncing and rolling objects. The Sound Design Toolkit (Monache, Polotti, & Rocchesso, 2010) contains an array of perception-based sound synthesizers that are designed to represent a variety of physical sound-producing events. Conan et al. (2013, 2014) developed a synthesis process in which sounds could be synthesized based on perceptually relevant action-object features. Their system allows auditory signal morphologies that define scraping, rubbing or rolling to be combined in continuous ways, allowing for a three-dimensional space of resemblance of generated sounds to each of those actions. Conan et al. (2014) noted that their framework could be used to create arbitrary action-object combinations. Pruvost et al. (2015) developed a sound synthesis system inspired by the work of Conan et al. (2014) and integrated it into a game engine, providing a way for XR developers to utilize synthesized action-object congruent sounds in XR applications. In the same way that visual physics simulation and procedural animation can augment hand-animation, sound production simulation could facilitate adherence to action-object congruency in XR environments.

1.4 Current Study

Prior to recommending the use of action-object congruent sound design and/or the use of procedural sound generation tools for XR applications, it was first necessary to evaluate whether action-object congruency is in fact impactful on MSI. The current study was designed to assess the contributions of action and object congruency to MSI in XR. A secondary goal was to investigate the interaction between these two types of congruency, and in particular to assess whether their combined effect could be described by a simple linear summation model.

CHAPTER 2. DEVELOPMENT OF STIMULI

2.1 Creation of Stimuli

Prior to conducting the study, a total of nine congruent auditory and visual stimuli were developed. These were comprised of three object types, and three depictions of the same three actions for each of those object types. Each action constituted a depiction of an *acting object* moving into contact with a, stationary *acted-upon object*. Both types of objects were identical in appearance and apparent physical properties, apart from the acted-upon objects being slightly larger.

Sounds and visuals were created with the following goals. First, the visual objects and sounds were designed to be novel and unrecognizable. Similarly, sounds were created via synthesizer rather than via modifying samples of real-world sounds. Second, parametric congruency features were held constant across stimuli. All object types appear to be the same size, and there are no differences in elevation or chroma. Third, visual stimuli were designed so that object and action properties would be readily discernable via brief visual assessment. Thus, stimuli were designed to minimize the possibility for semantic or parametric congruency effects to affect results, and to provide sufficiently clear visual and auditory information for action-object congruency to be determined by participants.

Sounds were generated using the sound design toolkit (SDT, Monache, Polotti, & Rocchesso, 2010), a free software suite that allows for synthesis of sounds using "audio algorithms which emphasize the role of sound as a process rather than a product." The SDT contains a variety of tools for synthesizing sounds based on physical events with certain

properties. Sounds produced by the SDT are intended to contain the features relevant to human perception of these different types of physical events, rather than being the result of exhaustive physical simulation. The SDT is comprised of a set of MAX MSP[2] patches. These tools were utilized to match properties of visual events to parameters within the SDT software. Sound clips generated in this manner were composited in Audacity[3] before being inserted into the Unity scene.

Four SDT tools were used: *impact, scraping, bubble* and *fluid flow.* For SDT Impact (Figure 1a), important parameters include the mass of the striking object ("hammer mass"), the velocity of the strike, and the stiffness of the object that is contacted. Three frequency components are synthesized, and can be adjusted. The decay of each frequency component can be individually adjusted, as can a global decay rate multiplier. Both settings affect the sustain of the synthesized sound. SDT Scraping (Figure 1b) models one rigid or semirigid surface being dragged across another. The "probe width" parameter refers to the amount of surface making contact between the two objects. This parameter is the key differentiator between whether a produced sound will tend to be perceived as a rub or a scrape, because it controls the magnitude and frequency of amplitude modulations (Conan et al., 2013). As with SDT Impact, there are three frequency components, a decay parameter for each, as well as a global decay multiplier. Finally, "velocity profile" can be modified to adjust the

[2] https://cycling74.com/

[3] https://www.audacityteam.org/

speed, duration and intensity of the scrape or rub. There is also a secondary "velocity" parameter. SDT Bubble (Figure 1c) is intended to synthesize the sound of a bubble forming and dissipating. Its key parameters are "bubble size," and "rise factor," the latter of which represents the extent to which the sound rises in frequency after onset. Finally, SDT Fluid Flow (Figure 1d) synthesizes many bubbles in at once or in rapid succession, and accordingly has a "bubbles per second" parameter that adjusts the rate of bubble production, along with parameters controlling the size range and rise range of generated bubbles.

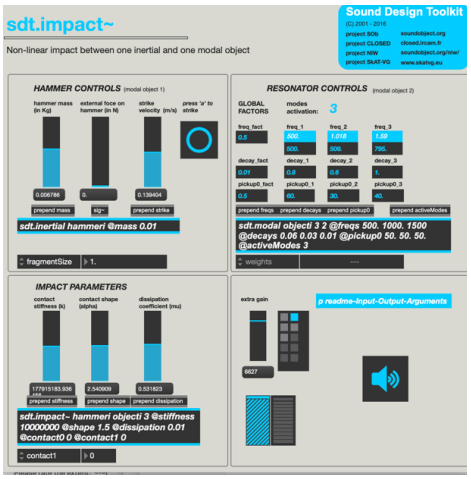

Figure 1a. SDT Impact user interface.

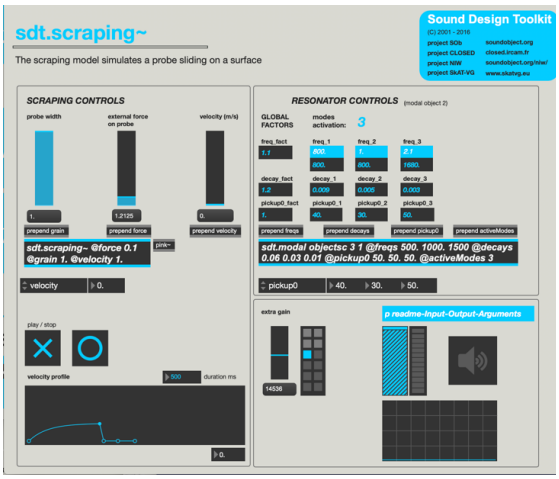

Figure 1b. SDT Scraping user interface.

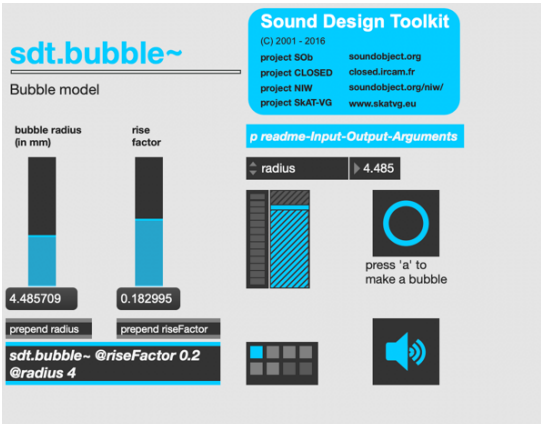

Figure 1c. SDT Bubble user interface.

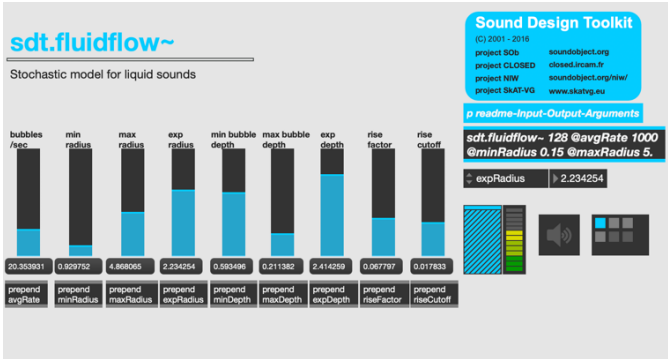

Figure 1d. SDT Fluid Flow user interface.

2.1.1 Smooth/Hard Objects

Smooth/hard objects were rendered as cubes, with a conical protrusion from one side. The objects had a normal map applied with low strength, generated using gaussian noise. This created the appearance of fine-grained, shallow surface irregularities, rather than the surface being entirely smooth. These objects were rendered with a moderate amount of specularity, and in cornflower blue.

2.1.1.1 Smooth/Hard Strike

For the visual event, the acting object was launched downward onto the acted-upon object at high velocity (Figure 2). The objects contacted evenly along a major face, and then movement immediately ceased.

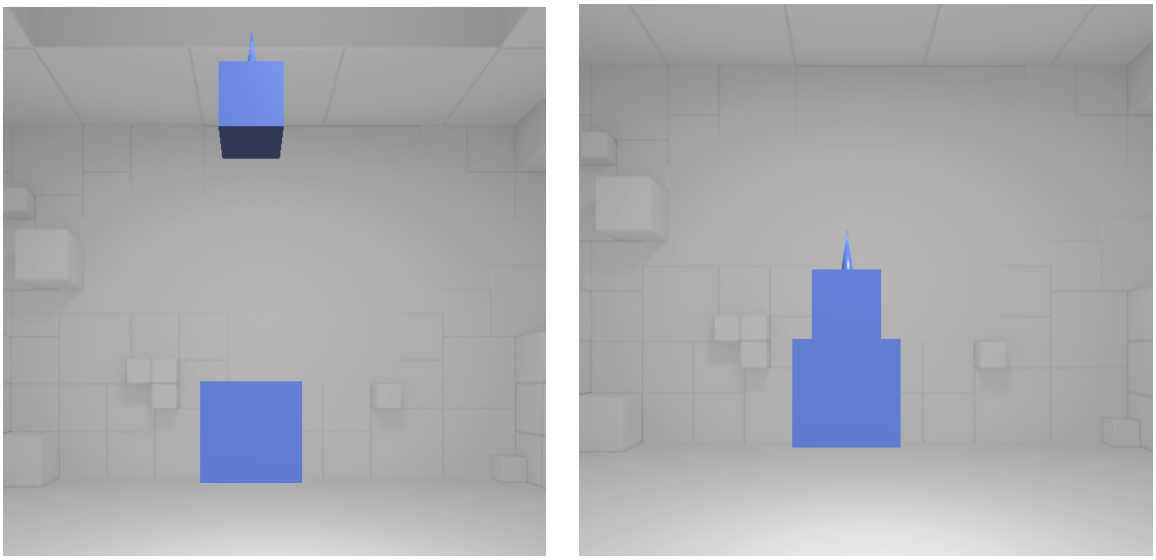

Figure 2. Smooth/hard strike visual event start (left) and end (right).

The sound was synthesized using the SDT impact synthesizer (Figure 3). The parameters were a hammer mass of .001 kg striking with a velocity of 3.74 m/s, with maximum contact stiffness. Frequency components were 800 Hz, 814 Hz, and 1590 Hz, with a decay factor of 0.25. This produced a staccato impact sound composed of primarily higher frequency components, with a brief higher-frequency sustained component.

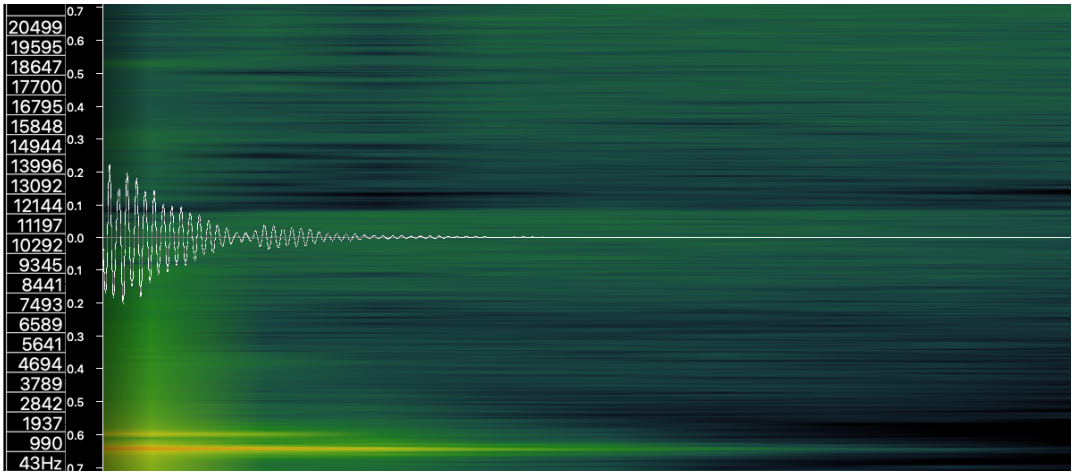

Figure 3. Smooth/hard strike waveform (inner ordinate/ white line, volts) and spectrogram (outer ordinate/ colored areas, dBV by frequency band in Hz).

2.1.1.2 Smooth/Hard Scrape

For the visual event, the acting object was pushed slowly across the top surface of the acted-upon object, over the course of approximately 550 ms (Figure 4). The end of the conical protrusion made contact with the upper surface of the acted-upon object, and moved across that surface. Slight random perturbations to the position of the acting object were applied as it moved.

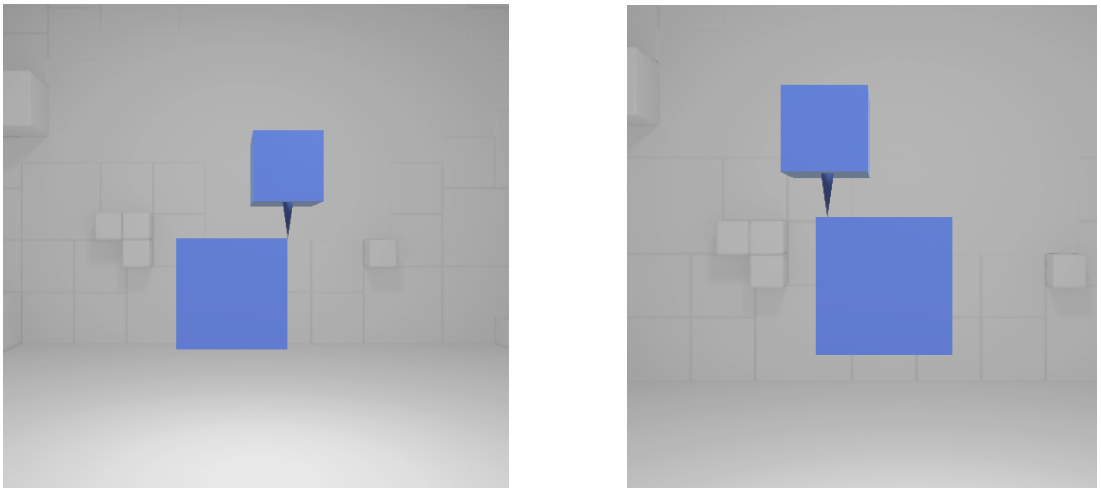

Figure 4. Smooth/hard scrape visual event start (left) and end (right).

SDT Scraping was used to synthesize the sound (Figure 5). Probe width was set to a low value of .0004, which produced a more irregular amplitude pattern indicative of scraping. Velocity was set to 0.96, and the velocity profile was slow and consistent over the course of approximately 550 ms. Frequency components were 4800 Hz, 7680 Hz, and 14,880 Hz. The decay factor was set to 1.5. This produced a relatively high frequency, inconsistent-amplitude sound, with minor higher frequency components indicating some amount of reverberation in the acted-upon object.

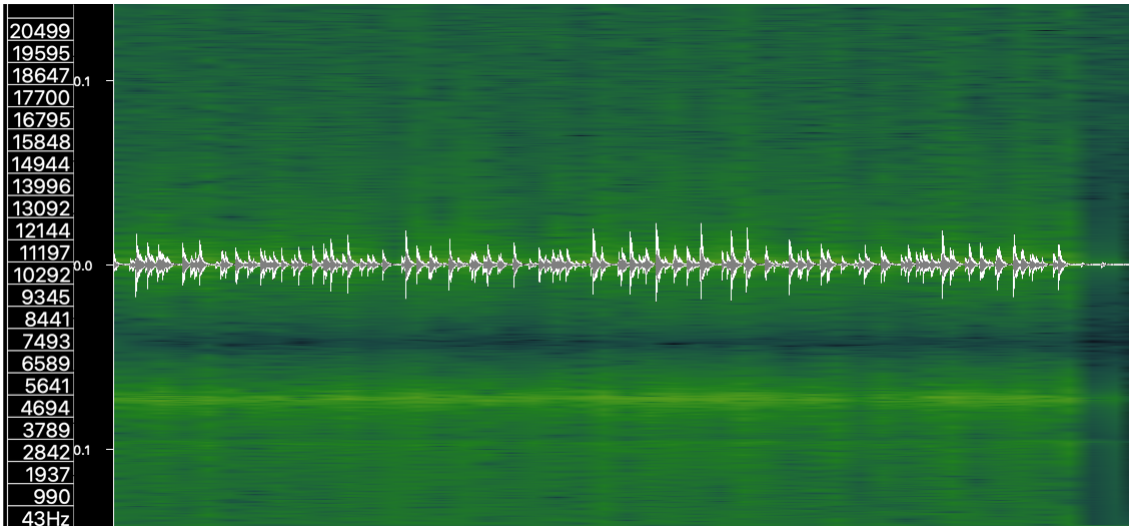

Figure 5. Smooth/hard scrape waveform (inner ordinate/ white line, volts) and spectrogram (outer ordinate/ colored areas, dBV by frequency band in Hz).

2.1.1.3 Smooth/Hard Rub

For the visual event, the acting object was pushed rapidly across the surface of the acted-upon object (Figure 6). Instead of the conical protrusion making contact, contact was made via a flat side of the acting object. Movement occurred over the course of approximately 200 ms.

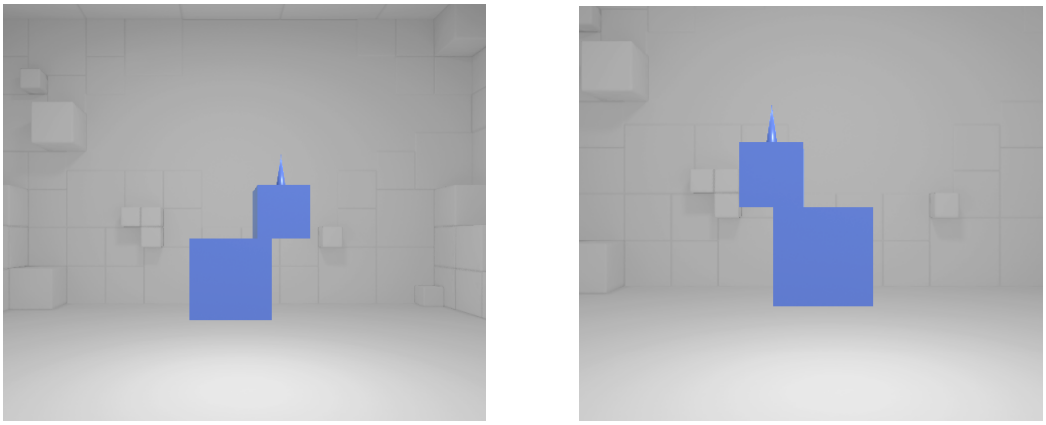

Figure 6. Smooth/hard rub start (left) and end (right).

For the sound (Figure 7), the SDT Scraping synthesizer was used. The maximum probe width (1) was used. Frequency components were 880 Hz, 880 Hz, and 1840 Hz. The velocity profile progressed with a brief onset from zero to moderate velocity, and returned rapidly to zero at 200 ms post-onset. The decay factor was set to 1.2, with additional frequency component decay parameters set to 0.009, 0.005, and 0.003. This produced a broadband rubbing sound, with extremely regular amplitude, and a low sustain.

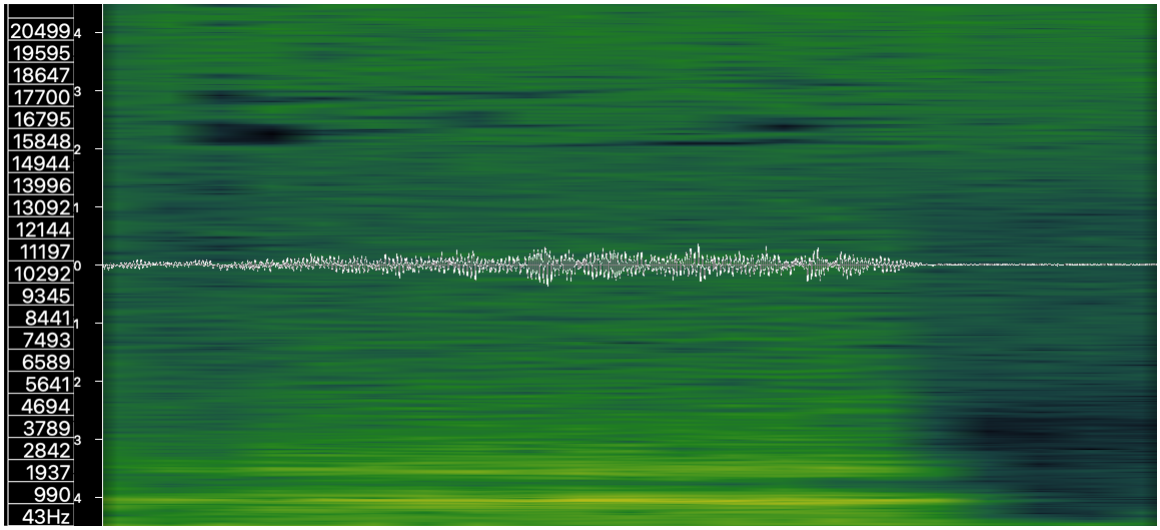

Figure 7. Smooth/hard rub waveform (inner ordinate/ white line, volts) and spectrogram (outer ordinate/ colored areas, dBV by frequency band in Hz).

2.1.2 Rough/Soft Objects

For the rough/soft objects, a cube with an uneven surface was created using random noise and tessellation. On top of this, a "fur" effect[4] was applied. This effect was tuned so that the "fur" was short to moderate in length, and was uneven enough to be perceptible given the resolution of the HMD and the distance that objects would be viewed at. The effect was also configured to not resemble any particular type of fabric, hair or fur, and to simply appear as a series of uneven semirigid follicular protrusions. The "fur" was rendered with a slight specular effect, a very limited metallic effect, and in cornflower blue.

2.1.2.1 Rough/Soft Strike

For the visual event, the acting object was launched at high velocity downward onto the acted-upon object (Figure 8). The acting object did not bounce.

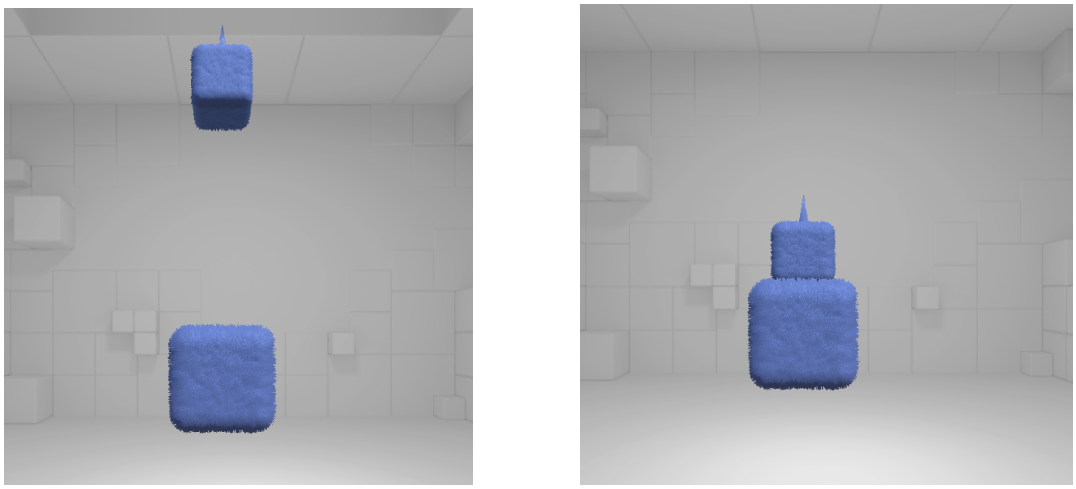

Figure 8. Rough/soft strike visual event start (left) and end (right).

[4] https://assetstore.unity.com/packages/vfx/shaders/imperial-fur-pbr-32522

The SDT Impact synthesizer was used. Hammer mass was .014 kg (moderately high), with a strike velocity of 0.14 m/s. Contact stiffness was moderate. Frequency components were 250 Hz, 254 Hz, and 398 Hz. The decay factor was 0.1. This produced a brief sound comprised of lower frequencies (Figure 9).

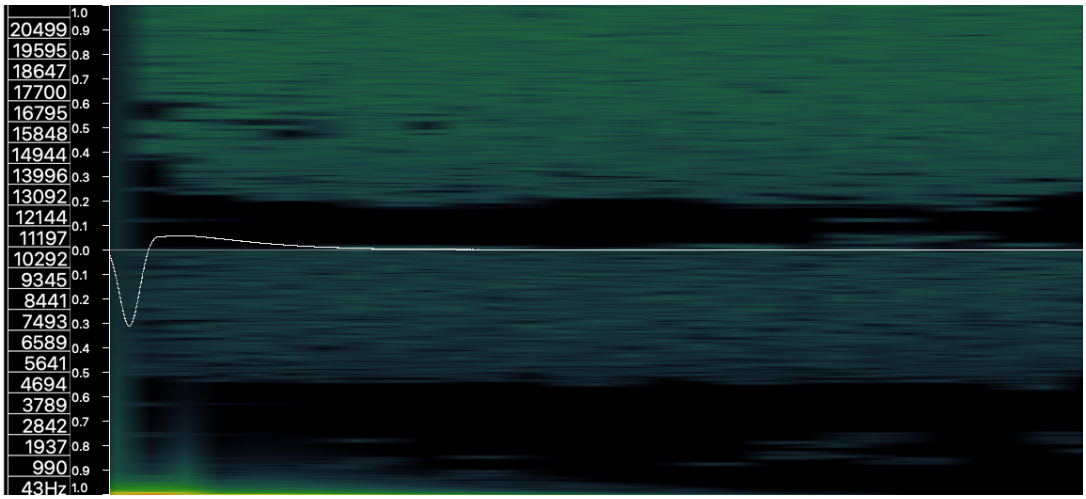

Figure 9. Rough/soft strike waveform (inner ordinate/ white line, volts) and spectrogram (outer ordinate/ colored areas, dBV by frequency band in Hz).

2.1.2.2 Rough/Soft Scrape

For the visual event, the acting object moved slowly across the top surface of the acted-upon object, over the course of approximately 550 ms (Figure 10). The acting object made contact with the acted-upon object only via the tip of the narrow protrusion. The tip of the protrusion was depicted as passing slightly into the layer of protrusions, implying flexibility on their part. Additionally, small random perturbations to the position were introduced as the acting object moved.

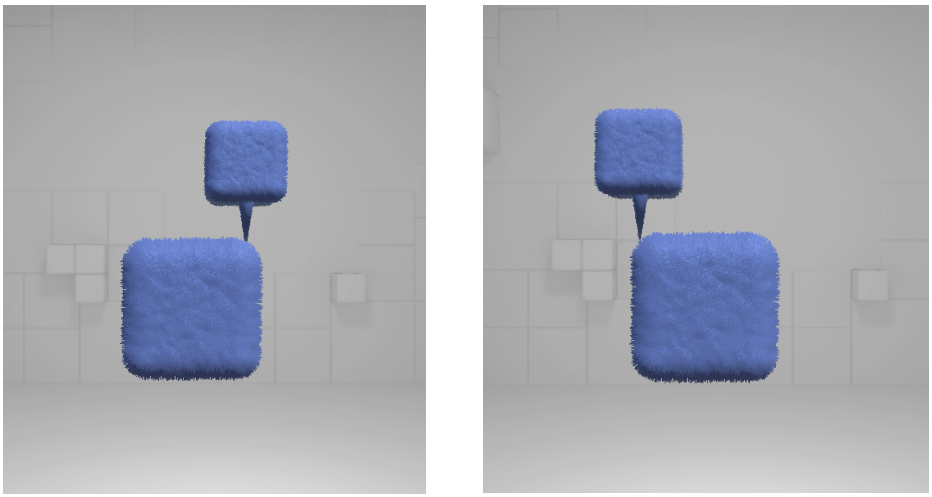

Figure 10. Rough/soft scrape visual event start (left) and end (right).

To produce the sound, the SDT Scraping synthesizer was used. The probe width was small (0.00008), and velocity high (3 m/s). Velocity profile rose to a moderate-high velocity, and dropped off rapidly at around 550 ms. Frequency components were 720 Hz, 1152 Hz, and 2232 Hz. The decay factor was set to 1, with low decay component values of 0.003, 0.002 and 0.002. This produced a sound with extremely variable amplitude, that was comprised of low to moderate frequencies, and had almost no sustain (Figure 11).

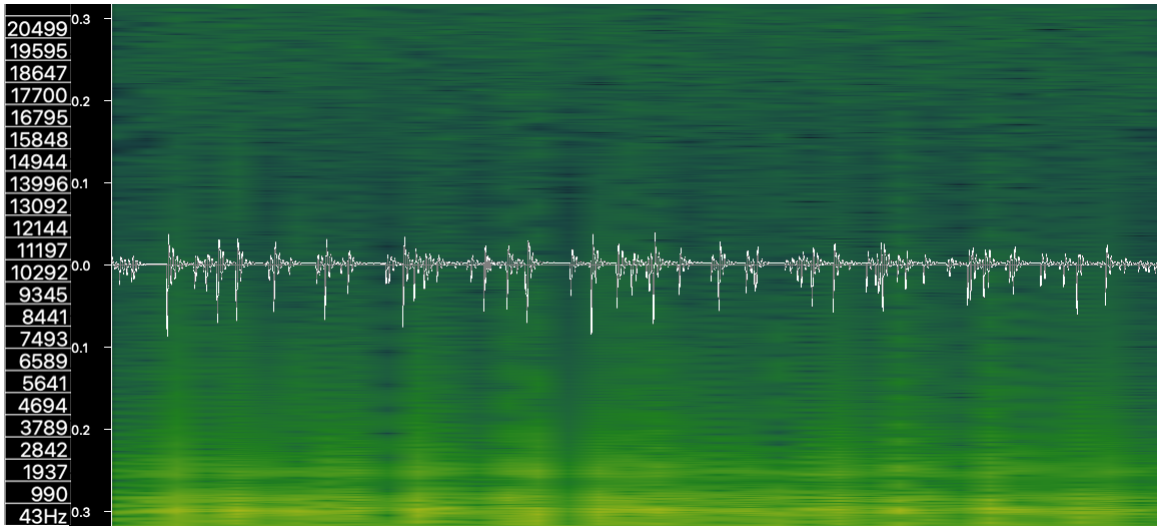

Figure 11. Rough/soft scrape waveform (inner ordinate/ white line, volts) and spectrogram (outer ordinate/ colored areas, dBV by frequency band in Hz).

2.1.2.3 Rough/Soft Rub

For the visual event, the acting object was moved laterally alongside the acted-upon object (Figure 12). Contact was made by the full bottom surface of the acting object. This motion took place over the course of approximately 200 ms.

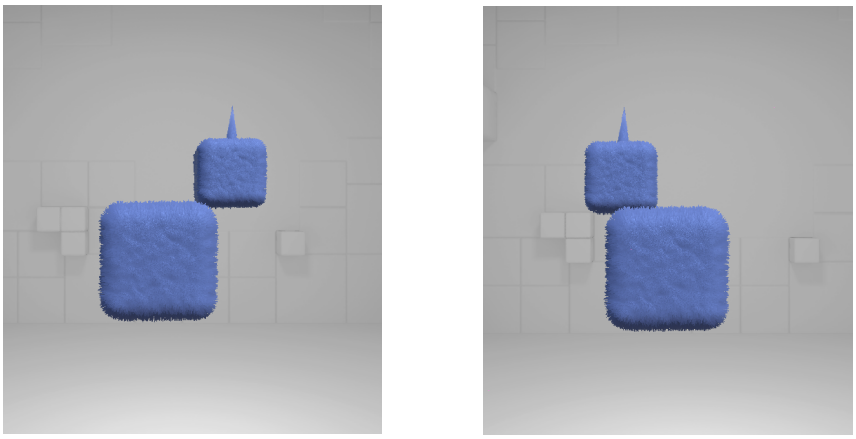

Figure 12. Rough/soft rub visual event start (left) and end (right).

SDT Scraping was used to generate the rough/soft rub sound. The probe width was moderate (0.005), and the velocity was very low (.005 m/s). Frequency components were 300 Hz, 487 Hz, and 930 Hz. The decay factor was set to 1, with decay components of 0.003, 0.002, and 0.002. The velocity profile rose quickly to a moderate velocity, before dropping to zero approximately 200 ms post-onset. This produced a broadband sound with moderate amplitude irregularities (Figure 13).

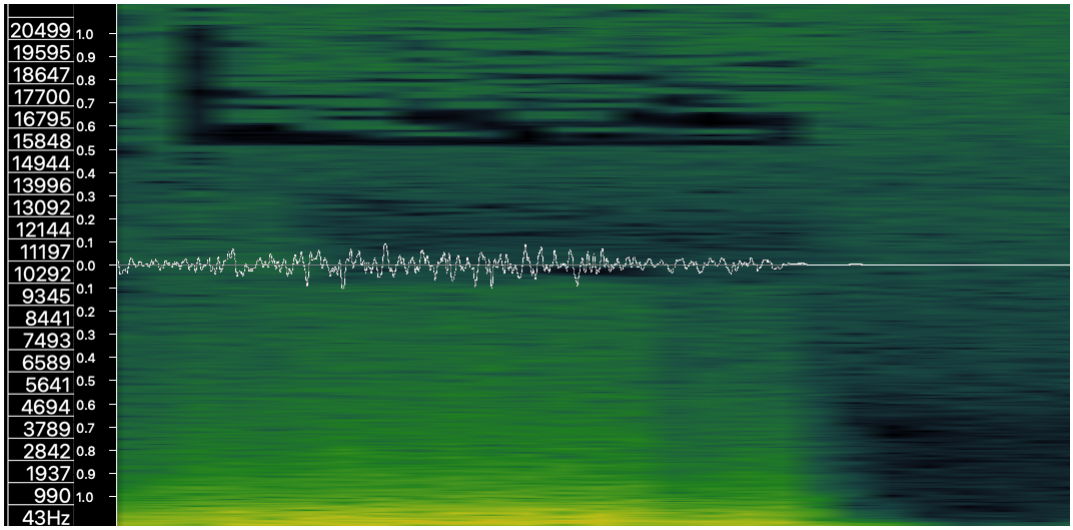

Figure 13. Rough/soft rub waveform (inner ordinate/ white line, volts) and spectrogram (outer ordinate/ colored areas, dBV by frequency band in Hz).

2.1.3 Gelatinous/Lumpy Objects

The gelatinous/lumpy objects were rendered as beveled cubes, with a seven by seven grid of protruding nodules on the top and bottom faces. One of these nodules protruded farther than the others, and was used for the scraping actions. These objects utilized softbody physics[5] simulation effects to render deformations that might occur as a result of movement or collision. They had high glossiness but low specularity, and were rendered in cornflower blue.

2.1.3.1 Gelatinous/Lumpy Strike

For the visual event, the acting object was launched with high velocity down onto the acted-upon object (Figure 14). The impact caused both objects to visibly deform, which was followed a moment later by a "rebound" motion as they returned to their original shape.

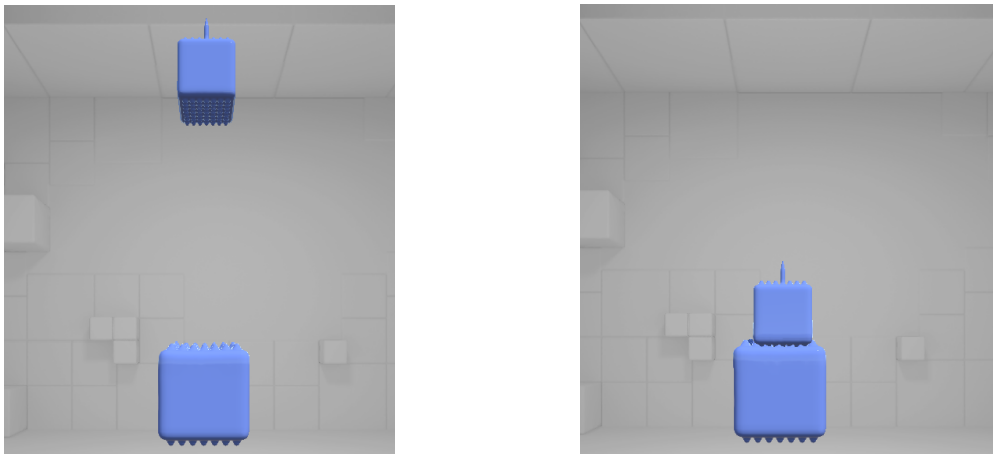

Figure 14. Gelatinous/lumpy strike visual event start (left) and end (right).

[5] https://assetstore.unity.com/packages/tools/physics/obi-softbody-130029

The sound was formed by compositing two SDT Bubble sounds. The first bubble component corresponded temporally with the initial impact and deformation, and the second corresponded with the "rebound," when the acted-upon object returned to its initial shape. The initial bubble was simulated with a radius of 4.48 mm (larger than other simulated bubble sounds), and a moderate rise factor of 0.18. The second bubble sound used the same settings but a different random seed, and was slightly increased in pitch/ decreased in amplitude (Figure 15).

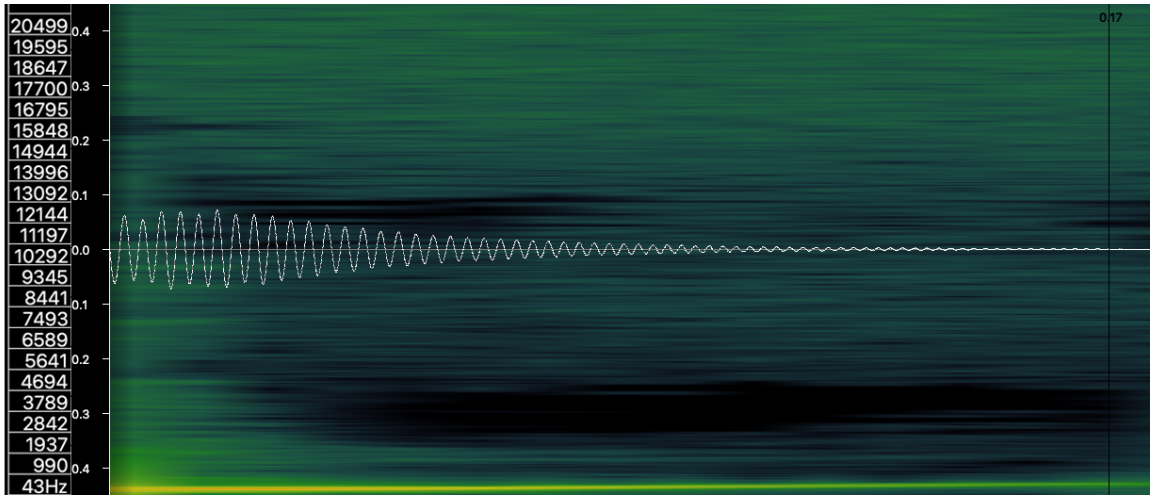

Figure 15. Gelatinous/lumpy strike waveform (inner ordinate/ white line, volts) and spectrogram (outer ordinate/ colored areas, dBV by frequency band in Hz).

2.1.3.2 Gelatinous/Lumpy Scrape

For the visual event, the acting object moved across the top surface of the acted-upon object, over the course of approximately 550 ms (Figure 16). The protrusion on the bottom of the acting object made contact with a series of seven of the nodules on the upper surface of the acted-upon object, which were seen to subsequently wobble and then return to their original shape.

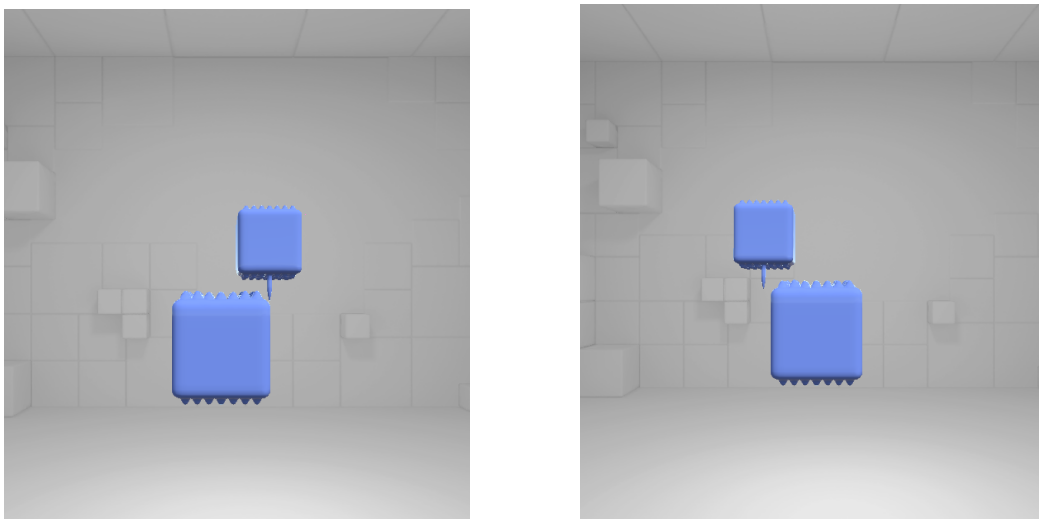

Figure 16. Gelatinous/lumpy scrape visual event start (left) and end (right).

The sound was a composite of seven SDT Bubble simulations, played in sequence. Each had a small bubble radius and modest rise factor, but these parameters were slightly different for each. The playback of the bubble sounds was timed so that they occurred in sync with each of the seven nodules being struck by the acting object (Figure 17).

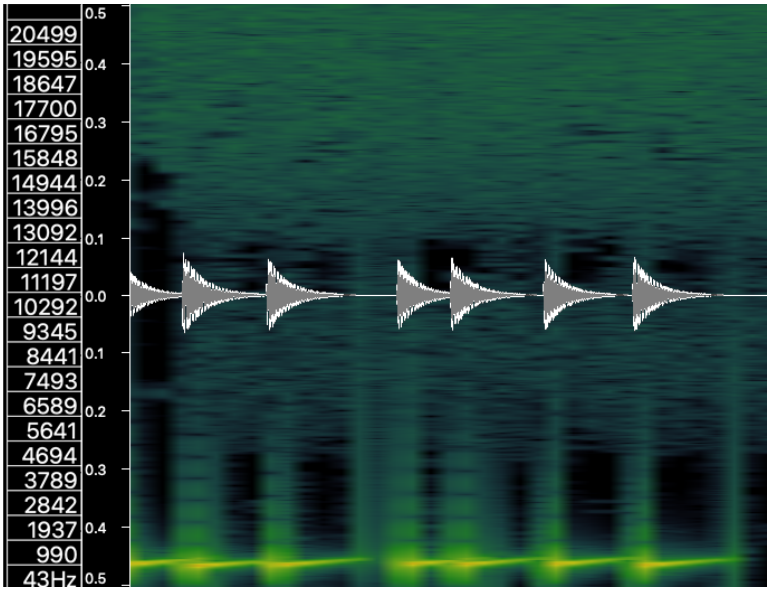

Figure 17. Gelatinous/lumpy scrape waveform (inner ordinate/ white line, volts) and spectrogram (outer ordinate/ colored areas, dBV by frequency band in Hz).

2.1.3.3 Gelatinous/Lumpy Rub

For the visual event, the acting object appeared to move across the top of the acted-upon object, causing significant deformation. The nodule grid on the bottom of the acting object made contact with the nodule grid on the top of the acted-upon object, leading to the visual depiction of many impacts. After the movement was completed, both objects returned to their original shape (Figure 18). This took place over approximately 200 ms.

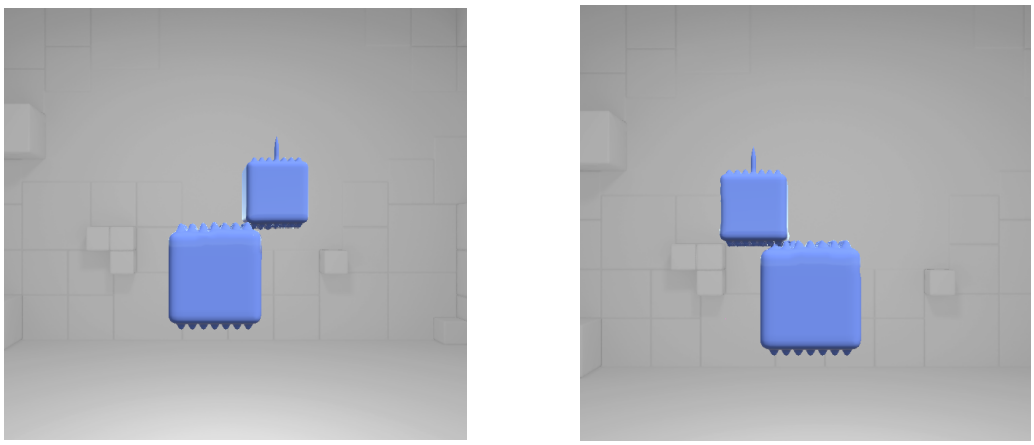

Figure 18. Gelatinous/lumpy rub visual event start (left) and end (right).

The sound was synthesized using the SDT Fluid Flow tool. Bubbles per second was set to 20, and other parameters were left close to their default values, allowing for a variety of bubble radii and rise factors. A clip of appropriate length was extracted, in which many such bubbles could be heard to form and dissipate. This paralleled the depiction of many collisions between gelatinous surface nodules, and subsequent returns to form, with the sound of a series of various bubbles being produced and subsequently "rising" (Figure 19).

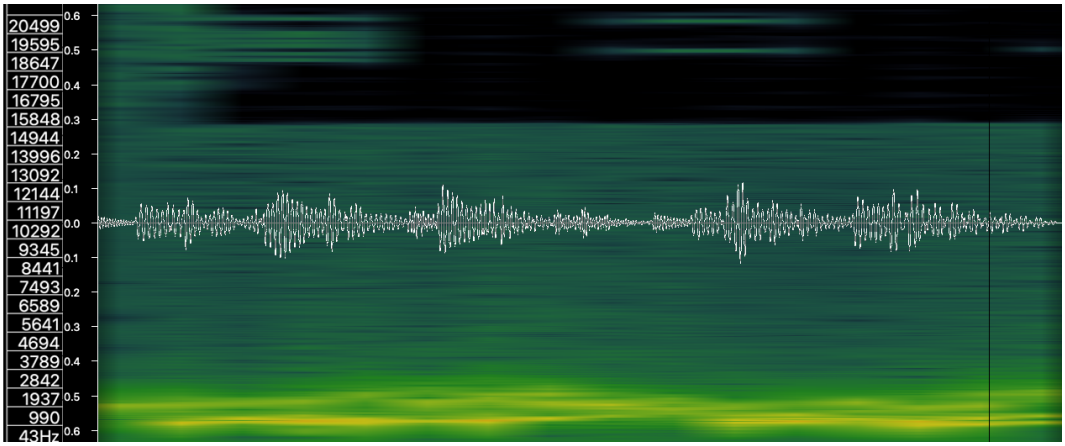

Figure 19. Gelatinous/lumpy rub waveform (inner ordinate/ white line, volts) and spectrogram (outer ordinate/ colored areas, dBV by frequency band in Hz).

2.2 Validation of Stimuli

2.2.1 Validation Survey Design

To validate stimuli, prior to conducting the laboratory sessions an online survey was administered (Appendix A). For each visual event, participants were asked to rate (via a Likert-type item) the extent to which each of the nine sounds matched that visual event, which was included as an embedded video. Videos were captured directly from the study software.

Additionally, participants were asked whether they perceived the stimuli depicted in the video to be novel, via two Likert-type items: (1) "The objects in the video remind me of other objects that I have seen, outside of this study," and (2) "It would be easy to name the material that the objects in the video are made of." Finally, participants were asked to name an object that the objects reminded them of, via a free-response question.

2.2.2 Validation Survey Results

Three rounds of this survey were completed utilizing a population of university undergraduates who were compensated with course credit for participating. The first round had 17 participants (7 male, 10 female, with a mean age of 22.6, SD = 6.73), the second had 16 participants (6 male, 10 female, with a mean age of 19.65, SD = 4.81), and the third had 25 participants (13 male, 12 female, with a mean age of 19.29, SD = 1.12). After the first and second iterations, descriptive statistics and grouping of free-responses were used to identify cases in which sounds either (a) were not sufficiently well matched

with their intended visual event, or (b) were identifiable and/or commonly associated with specific material or type of object.

After the final iteration, the majority of sounds were rated as best matched to their intended action and objects (via descriptive statistics). There were two cases in which this was not true. For the hard/smooth rub sound, the soft/rough rub animation was rated as matching slightly better than the intended animation. Similarly, the soft/rough scrape animation was rated as matching slightly better to the hard/smooth scrape sound than the intended animation. The final survey iteration also indicated that the visual objects were generally novel and unfamiliar, with responses averaging below the midpoint for both questions assessing stimulus novelty.

CHAPTER 3. METHOD

3.1 Participants

After completing the stimulus validation process, the main study was conducted with 28 university undergraduates. There were no exclusion criteria other than participants having normal or corrected-to-normal vision and hearing, and sufficient dexterity to respond during trials. Participants were compensated with 1 hour of course credit for up to 1 hour of participation in a single session.

The average age of the participants was 19 (*SD* = 1.20). Twenty-one participants identified as male and 7 participants identified as female. Participants scored an average of 53.10 on the GOLD-MSI musical sophistication index (*SD* = 13.94). This indicates that the sample had moderate musical sophistication. The majority of participants had limited exposure (less than 5 hours) to VR. No participants reported having over 24 hours of exposure to VR, and only three participants reported owning a VR headset.

3.2 Apparatus and Materials

3.2.1 Physical Environment and Audio Hardware

The experiment took place inside a small room, with sufficient space for the participant and a set of speakers (Figure 20). An adjustable-height drafting chair was located in the centre of the room. Surrounding this chair was a circle of seven Eris E5 speakers. These speakers were studio monitors, characterized by flat, neutral response[6].

[6] https://www.presonus.com/products/eris-e5

Audio output was routed out of the control computer through an audio interface, providing 7.0 surround output to the speakers. The control computer was placed just outside the room to minimize audible fan noise.

Figure 20. Physical study environment and apparatus.

Audio output from the control computer was tuned to minimize playback latency, while avoiding the clicking and popping symptomatic of extremely low-latency playback. Tuning the software and hardware in this manner also reduced latency *variability* in addition to reducing average latency. However, since even a brief latency could have impacted results, the following procedure was carried out to compensate for latency. First, rather than sounds being triggered to play at the moment of contact between the acting and acted-upon objects, the process of playing the sound was initiated 31 ms prior to the visual depiction of the action occurring. The 31 ms value was arrived at through recording video and audio with a 120-fps slow-motion camera and microphone, assessing whether the sound onset and visual event appeared to occur on the same frame, and adjusting the magnitude of the temporal offset value. It should be noted that this offset value was specific to the hardware and software used in this study.

The registration of the virtual sound space and the real speakers was calibrated using an iterative adjustment procedure (Figure 21). First, an electronic protractor with attached pointer was used to measure the angle of each speaker relative to the chair (specifically the rear-center of the chair, where a participant's head would be). These values were used to construct virtual speakers in the Unity scene. A sound was then played from one of these speakers (white noise). An evaluator pointed in the direction they judged that sound to be coming from, and the experimenter moved the 3D speaker model (via rotation about the center of the scene) until it matched the direction in which they were pointing. Using the protractor, the real-world speakers were then adjusted to match the new angle. This procedure was repeated until subsequent evaluators were no longer requiring changes in position in order for perceived sound source location and virtual sound source location to by in sync. Documentation of Unity's implementation of 7.1 surround was not available, and that the 7.1 specifications[7] do not specify the precise positions of speakers. As such, the aforementioned procedure was necessary.

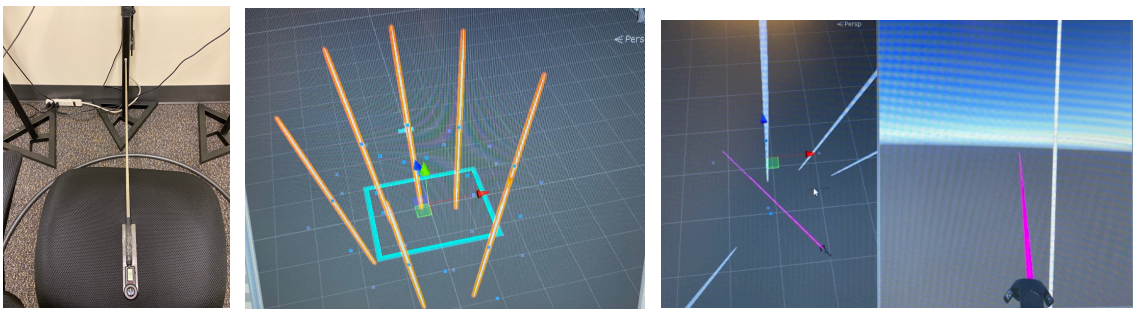

Figure 21. Electronic protractor (left), virtual reproduction of speaker locations (center), and an expert pointing toward a localized sound (right).

[7] https://www.dolby.com/us/en/guide/surround-sound-speaker-setup/7-1-setup.html

3.2.2 Virtual Environment

The VE was constructed using the Unity real-time development platform[8], and the VR toolkit SteamVR[9]. These tools interfaced with an HTC Vive Pro headset[10] to provide visual output, as well as the ability for the study software to receive input from the handheld Vive Pro controllers (Figure 23). The Vive Pro controllers were tracked, in addition to the HMD, which allowed for precise pointing by participants. Positional drift and noise were minimal with this setup, which utilized version 2.0 outside-in lighthouse tracking[11]. Luckett (2018) found that, with the older 1.0 tracking system, tracked positions of the HMD and controllers were not significantly different from those produced by a high-precision laser-tracking system, as long as tracking was not lost entirely.

Figure 22. HTC Vive Pro HMD, lighthouse trackers, and controllers.

[8] https://unity.com/

[9] https://store.steampowered.com/app/250820/SteamVR/

[10] https://www.vive.com/us/product/vive-pro/

[11] https://www.valvesoftware.com/en/index/base-stations

The VE was rendered using a sufficiently powerful control computer capable of rendering frames quickly enough to refresh the display at the maximum possible rate (90 Hz) while still maintaining the full resolution of the Vive Pro (1400x1600 per eye). This, combined with the objects being relatively large and located relatively close to the participant, allowed the texture of the objects to be perceptible.

The VE itself (Figure 23) was a grey room with a similar shape and slightly larger size compared to the physical experiment environment. In the early stages of study design, featureless spaces were utilized. Pilot participants indicated that this was disconcerting, so the final VE was designed to include some basic features (wall cubes, a ceiling light fixture, and other geometry), while maintaining the neutrality and freedom from distraction typical of a screen-based perception study. The VE was also changed from white to grey in response to pilot sessions in which participants indicated that they had experienced eye fatigue.

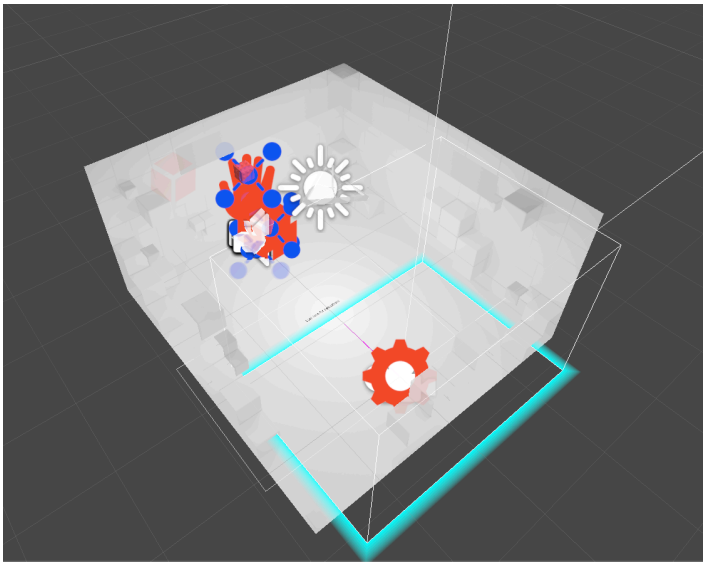

Figure 23. Overview of the VE. The participant's viewpoint is represented in the lower right as a gear icon/camera icon. Icons in the upper left indicate the locations of stimuli.

3.2.3 Sound Rendering

Limited room reverberation, reflection and echo effects were utilized to simulate sound propagation within the virtual space, and facilitate binding of sound played over world-space speakers to virtual-space virtual objects. The parameters were set as shown in Figure 24.

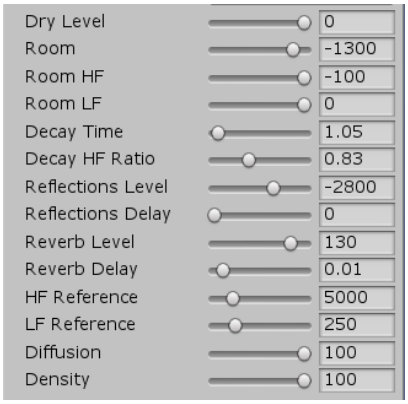

Figure 24. Reverberation, reflection and echo settings.

These settings produced a moderate reverberation, appropriate to the apparent size of the VE, and minimal reflections and echoes. The reverberation, reflection and echo effects were tuned to: (a) facilitate binding of sounds played over the speakers to virtual object interactions, and (b) to prevent the occurrence of outright confusion as to sound source location and timing. These goals led to the use of moderate reverberation and minimal reflections and echoes (to prevent front-back reversals from occurring due to rear-rendered reflections).

The loudness of sound playback was tuned in a similar manner. First, sounds needed to be loud enough to be clearly perceptible and localizable, while still having some variation in sound loudness that would be expected given the parameters of the depicted

visual action. This meant that some sounds, such as the Smooth/Hard Rub sound, were adjusted to be louder than they might have been in the real-world, relative to sounds such as the Smooth/Hard Strike sound.

3.3 Procedure

3.3.1 Device Fitting and Calibration

After consenting to participate, participants sat on the drafting chair in the middle of the circle of speakers. The wheels of the chair had been removed, and the seat locked from reclining, so that the chair was unable to move, aside from rotating. They were instructed to sit comfortably in the chair, and to adjust the height of the chair so that a piece of tape just above the center speaker was at their eye level. This procedure placed the head of the participant in the center of the speaker ring, horizontally as well as vertically. Finally, participants were instructed not to slouch to the left or right during the course of the study.

After calibrating the seat height, the participant was given the HMD and instructed in how to fit and adjust the device. After they put it on, the experimenter asked if they could see clearly, or if there were blurry/glowing visuals symptomatic of an incorrect fit. If the fit was incorrect, the experimenter assisted with ensuring a good fit. The built-in headphones were not placed over the participant's ears, since sounds were to be played over the speakers instead of these headphones. After the participant was seated and fitted with the headset, the experimenter gave them a VR controller and then launched the experiment software.

3.3.2 Spatial Ventriloquism Task Training Phase

Before trials began, an automated training procedure familiarized participants with the spatial ventriloquism task. The training procedure had three phases. The training procedure introduced the sounds first, and then the visual events, in order to prevent a response strategy that was observed in some early pilot sessions in which participants always pointed toward the visual event. Throughout the study, participants heard instructions spoken over the speakers, and subsequently had the option to read a written version within the VE. As shown in Figure 25, a text version of instructions remained visible throughout all trials, positioned at the participant's feet.

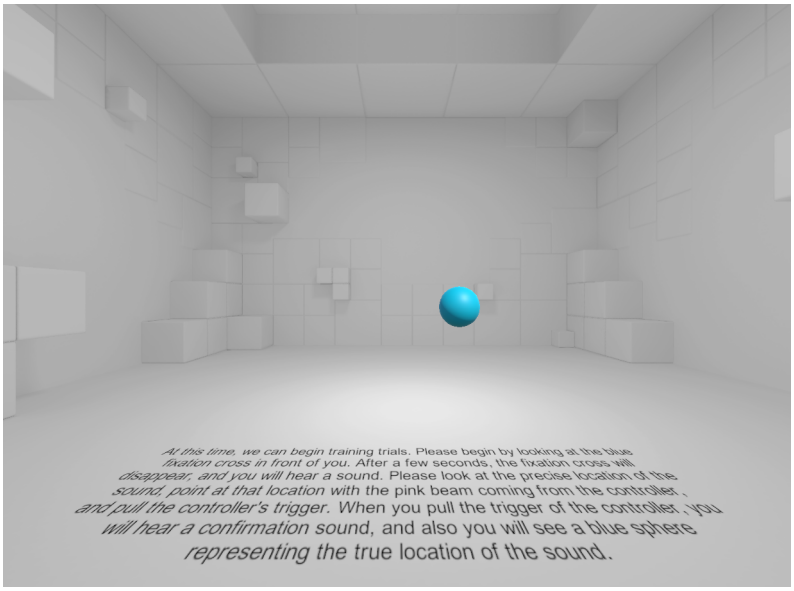

Figure 25. Blue sphere confirming the location of the sound during the first training phase, and onscreen instructions.

In the first phase, participants started by viewing a fixation cross. Then, a sound played from one of the speakers. Participants were instructed as follows: "Please look at the precise location of the sound, point at that location with the pink beam coming from the controller, and pull the controller's trigger." After pulling the trigger on the controller,

a blue sphere was presented to participants to indicate the true location of the sound. In this first phase of the training, there were no interacting visual objects present. The training procedure started with this phase in order to familiarize participants with the response method of pointing toward sounds that they heard.

In the second phase of spatial ventriloquism task training, participants also saw visual events occurring when the sound played in each trial. They were instructed as follows:

> "During the next few trials, you will see two objects interacting and hear a sound. Sometimes the sound will come from the same location as the objects, and sometimes it will not. If you think that the sound and the interaction between the objects occurred at the same location in space, point the pink beam directly at the objects and pull the trigger. If you think they did not occur at the same location in space, please point to the location of the sound."

This phrasing was designed to suggest that the sound could come from the interacting objects, or not, in response to some pilot participants who chose to response strategically by always pointing toward the objects or always pointing away from the objects. Participants continued to be able to see the blue sphere after they responded, and thus remained able to assess how their response compared to the true sound location.

Finally, in the third phase of the training, participants were given the same instructions, but were advised that they would no longer see the blue sphere after each trial (and thus would not have confirmation of the true location of the sound).

Participants completed 9 trials in the first training phase, 5 trials in the second phase, and 4 trials in the third phase. After they completed all of these trials, training was complete, and the experimenter paused the study software.

3.3.1 Administration of Questionnaire

Since participants by this point had been within the VE for a few minutes, they were then instructed to remove the HMD and to fill out the simulator sickness questionnaire using a tablet (SSQ; Kennedy, Lane, Berbaum, & Lilienthal, 1993; see Appendix B). The software automatically computed a result based on the responses of the participant. If the participant responded in a manner indicating they may have been experiencing simulator sickness (operationalized as a composite score of 2 or greater, with 3 being the maximum possible score and 0 indicating no symptoms), the software recommended that they did not proceed. This did not occur for any study participant.

Once the SSQ indicated that the participant was not experiencing simulator sickness, the tablet automatically advanced to a set of basic demographic questions, several questions that assed whether the participant was able to see and hear the stimuli, and two questions about VR experience (Appendix D). These were followed by the Goldsmith Musical Sophistication Index (GOLD-MSI), which assesses musical sophistication (Müllensiefen, Gingras, Musil, & Stewart, 2014; Appendix C).

After completing the questionnaire, participants were asked to put the HMD back on, and retrieve the controller. When they indicated they were ready to continue, the experimenter un-paused the software, and the study proceeded.

3.3.2 Spatial Ventriloquism Trial Structure

Next, participants experienced two blocks of 146 spatial ventriloquism trials of the sort they had practiced during the last phase of training. Spatial ventriloquism tasks were recommended by Bruns and Röder (2019) as a way of assessing MSI strength, in particular due to their ability to produce a continuous measure. Each trial proceeded in the following manner.

First, the software checked if the HMD was facing forward. If it was not, the next trial would not be administered. If it was, a fixation cross was presented centrally (as shown in Figure 26) for a randomly determined interval of 400-700 ms. After this interval, the two objects appeared. Before the acting object started to move, both objects were present for an interval of 600 ms, to allow the participant to briefly visually assess them.

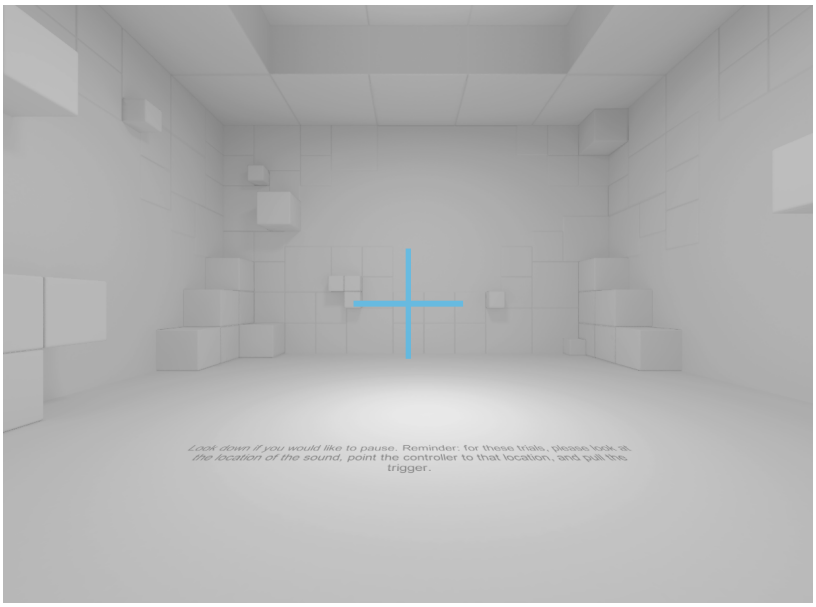

Figure 26. Participant view during fixation period, with fixation cross and instructions.

After this interval, the acting object began to move, resulting in an interaction with the stationary acted-upon object. The acted-upon object was always located centrally along the azimuth (in the location where the fixation cross had been previously), and was elevated so that the top face of the acted-upon object was aligned with the horizontal bar of the fixation cross. This allowed the action to occur centrally.

When the two objects came into contact, this constituted one of the three actions, or the beginning of such (striking, scraping or rubbing). A sound played when the objects came into contact. This sound was offset along the azimuth by either 6 or 12 degrees.

Acting and acted-upon objects were rendered at a distance of 1.5 meters from the participant. The position of these objects (as well as the fixation cross) was continually centered on the HMD. This was done so that, if the participant moved (translated) their head slightly, objects would still appear with the desired positioning relative to the HMD. See Figure 27 for an example participant view of a visual event.

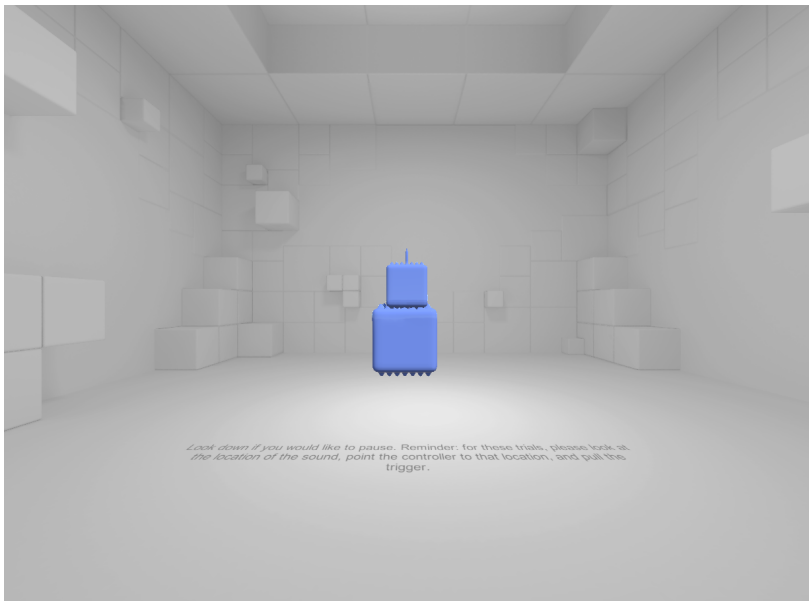

Figure 27. Participant view with visual objects and instructions.

After the sound played, participants were tasked with looking at, then pointing toward, the location of the sound, or the location of the objects if they thought the sound came from the objects (see section 3.3.2 for the exact wording).

Rather than the participant being able to respond immediately, there was an interval of 1300 ms in which the pointing beam was not visible, and the participant was unable to respond. This interval was included to prevent the reflexive or casual responding that was occasionally observed during pilot sessions, and to instead encourage more deliberate and precise responses.

When the participant pulled the trigger on the VR controller, the objects disappeared, and the fixation cross reappeared, beginning the trial sequence for the subsequent trial. See Figure 28 for the full trial sequence.

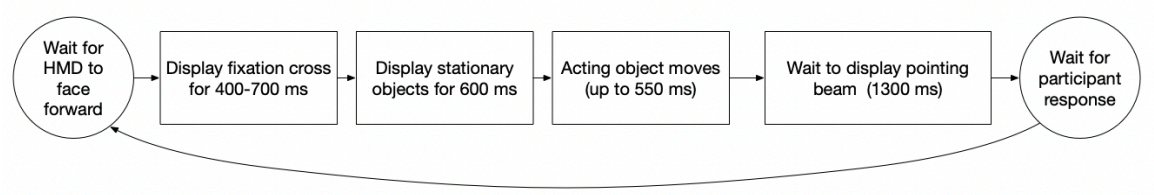

Figure 28. Spatial ventriloquism trial sequence.

These spatial ventriloquism trials produced a *localization biasing* score. This score was calculated by the software as the rotational distance (in degrees of visual angle) between the true direction of the sound and the direction of the participant's response. Although participants could respond by pointing above or below the azimuth, the software only utilized the component of their response that corresponded to the azimuthal coordinate, since sounds and visual stimuli were always presented along the azimuth. Negative localization biasing values indicated that a participant responded by pointing in

a direction opposite the direction of the visual event. These negative values were not removed. Thus, if localization biasing scores were positive on average, this indicated that spatial ventriloquism was taking place. If no biasing were taking place, the expected value of these scores would be zero, reflecting unbiased localization error. Scores could also be negative on average, reflecting a phenomenon in which stimuli that tend *not* be bound can cause systematic "reverse" localization biasing (Wallace et al., 2004b).

During the course of these trials, participants experienced every possible combination of Action-Incongruent-Object-Incongruent, Action-Congruent-Object-Incongruent, Action-Incongruent-Object-Congruent, and Action-Congruent-Object-Congruent stimuli, for each of the two audiovisual offsets. In the two partially-congruent conditions, each offset/stimuli combination was presented twice, and in the fully congruent conditions it was presented four times, to compensate for the differing number of combinations in each condition. This produced a similar, but not identical, number of trials in each of the four experimental conditions.

Once trials were generated, they were administered in a randomized order. There was a built-in break period halfway through the spatial ventriloquism trails. When all spatial ventriloquism trials were completed, there was an additional break period.

3.3.3 Temporal Ventriloquism Trial Structure

After the second break period, participants began the second part of the procedure, in which they completed a set of 166 temporal ventriloquism trials. These trials were designed to assess a different aspect of MSI strength: the *perception of unity* (Bruns & Röder, 2019). Although perception of unity tends to be correlated with the amount of

localization biasing (Wallace et al., 2004b), response biasing and perception of unity reflect two different aspects of MSI (Chen & Spence, 2007). As such, these were measured via two separate types of trials, and were analyzed using univariate statistical methods.

For these trials, participants were instructed to pull the trigger on the controller if the visual event and sound occurred at the same time, or do nothing if they occurred at different times. A "different" response was recorded if the participant did not pull the controller's trigger within 2.75 seconds. In temporal ventriloquism trials, sounds were played with SOAs of either 170 ms or 180 ms. Participants were not trained in how to complete the temporal ventriloquism task, which was simple to explain and carry out. The between-trial sequence of events was the same for temporal ventriloquism trials as it was for spatial ventriloquism trials. See Figure 29 for the full trial sequence.

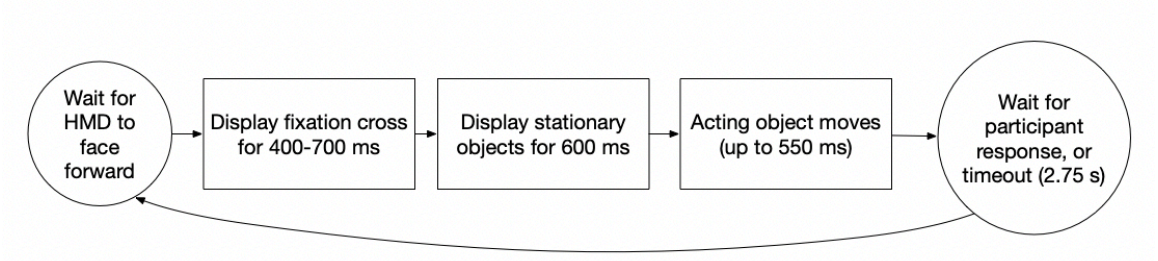

Figure 29. Temporal ventriloquism trial sequence.

For temporal ventriloquism trials, a *simultaneity judgment rate* variable was produced. This variable was the rate at which the participant judged the auditory and visual stimuli to have occurred at the same time. This was *not* an accuracy variable, since the two stimuli were in fact never presented at the same time.

As before, participants experienced every possible combination of Action-Incongruent-Object-Incongruent, Action-Congruent-Object-Incongruent, Action-

Incongruent-Object-Congruent, and Action-Congruent-Object-Congruent trials, for each of the two SOAs. Temporal ventriloquism trials were administered in an entirely random order. After completing all temporal ventriloquism trials, the study procedures were complete. Participation took up to one hour.

3.4 Research Design

3.4.1 Experiment Conditions

Each participant experienced all four conditions (see Table 1). Rather than experiencing these conditions in discrete blocks, each trial could be in any of the four conditions, with the order of trial administration determined randomly.

Table 1 – Conditions experienced by each participant.

	Object-Incongruent	**Object-Congruent**
Action-Incongruent	Action-Incongruent-Object-Incongruent	Action-Incongruent-Object-Congruent
Action-Congruent	Action-Congruent-Object-Incongruent	Action-Congruent-Object-Congruent

3.4.2 Analyses

First, Hyunh-Feldt two-way repeated measured ANOVAs were conducted for both dependent variables. Next, three sets of planned paired t-tests were conducted. The first compared the Action-Congruent-Object-Incongruent to the Action-Incongruent-Object-Congruent condition, and the second compared the Action-Incongruent-Object-Incongruent condition to the Action-Congruent-Object-Congruent condition. These tests were conducted both for localization biasing scores and simultaneity judgment rates.

Prior to the final planned t-test, for each participant, difference scores were calculated for the Action-Congruent-Object-Incongruent, Action-Incongruent-Object-Congruent, and Action-Congruent-Object-Congruent conditions by subtracting the localization biasing score for the Action-Incongruent-Object-Incongruent condition from each. The difference scores for the two partially-congruent conditions were then summed to create a *superadditivity threshold*. Then, a paired t-test was conducted comparing the superadditivity threshold to the participant's Action-Congruent-Object-Congruent difference score.

For the five aforementioned planned t-tests, a family-wise alpha of .05 was maintained by applying Bonferroni corrections.

Before conducting any analyses, two participants who exhibited negative localization biasing for all four conditions were removed. This condition was selected a priori in response to the observation that some pilot participants did not appear to experience the ventriloquism phenomenon. This could have been due to the fact that all

participants experienced the same spatial offsets, even though individuals vary in sound source localization ability and sensitivity to spatial co-locatedness, or to some other cause.

3.4.3 Hypotheses

It was expected that adhering to action congruency and object congruency, individually, would lead to increased localization biasing on spatial ventriloquism trials and increased simultaneity judgment rates on temporal ventriloquism trials compared to fully incongruent trials, and that fully congruent stimuli would lead to larger differences in the two dependent variables than partially congruent stimuli (hypothesis 1). Of those two partially congruent conditions, Action-Congruent-Object-Incongruent stimuli were expected to lead to significantly greater localization biasing and simultaneity judgment rates compared to Action-Incongruent-Object-Congruent trials (hypothesis 2). Finally, it was hypothesized that localization biasing scores in the fully congruent condition would be significantly greater than the superadditivity threshold (hypothesis 3).

CHAPTER 4. RESULTS

4.1 Hypothesis 1: Validation of Action-Object Congruency

As shown in Figure 30, the planned paired t-test indicated that the Action-Congruent-Object-Congruent condition led to greater localization biasing (M = 2.77, SD = 1.99) compared to the Action-Incongruent-Object-Incongruent condition (M = 1.61, SD = 1.92), $t(27) = 4.469$, $p < .001$, $d = 1.54$.

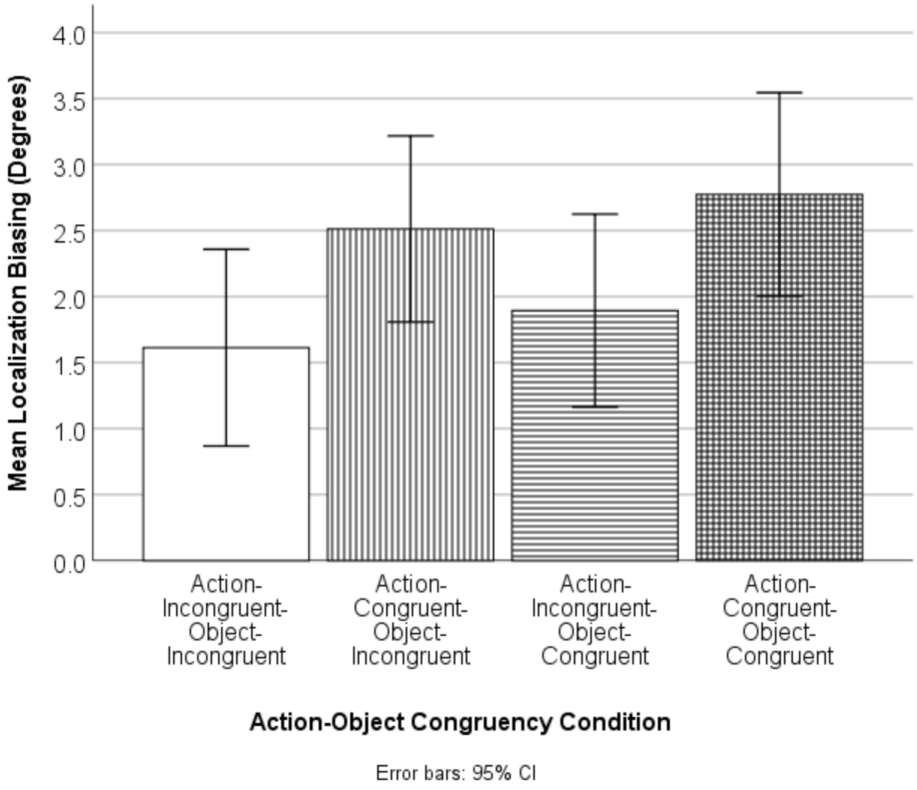

Figure 30. Mean localization biasing by action-object congruency condition.

As shown in Figure 31, the Action-Congruent-Object-Congruent condition also led to higher simultaneity judgment rates (M = 0.83, SD = 0.16) compared to the Action-Incongruent-Object-Incongruent condition (M = 0.63, SD = 0.16), $t(27)$ = -5.135, $p < .001$, d = 1.27.

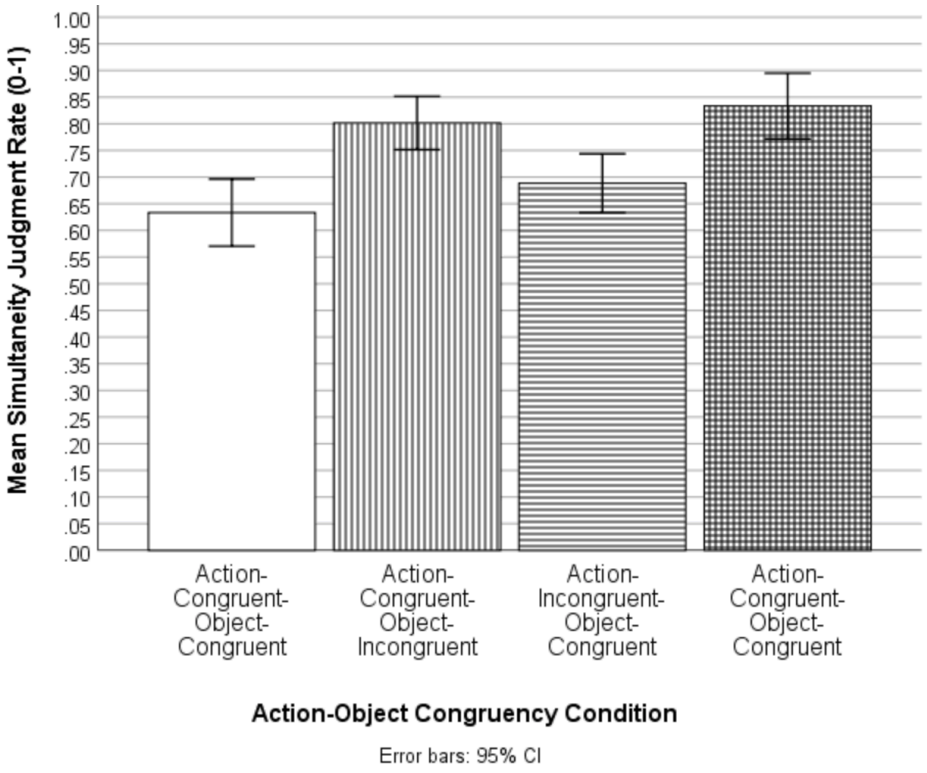

Figure 31. Mean simultaneity judgment rate by action-object congruency condition.

4.1.1 Action Congruency

There was a significant main effect of action congruency on localization biasing, $F(1,27) = 22.353$, $p < .001$, $\eta^2 = .453$. Participants exhibited greater localization biasing when presented with action-congruent stimuli (M = 2.64, SD = 1.84) compared to action-incongruent stimuli (M = 1.75, SD = 1.86).

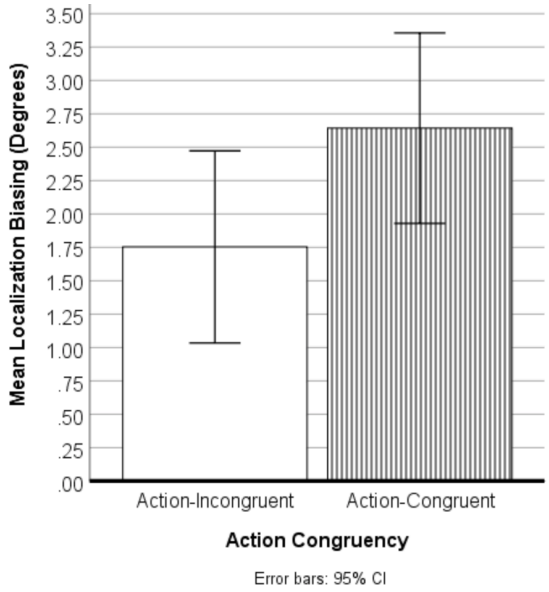

Figure 32. Mean localization biasing by action congruency.

There was also a significant main effect of action congruency on simultaneity judgment rate, $F(1,27) = 35.424$, $p < .001$, $\eta^2 = .567$. Participants indicated that stimuli were simultaneous at a higher rate when presented with action-congruent stimuli ($M = 0.82$, $SD = 0.13$), compared to action-incongruent stimuli ($M = 0.66$, $SD = 0.14$).

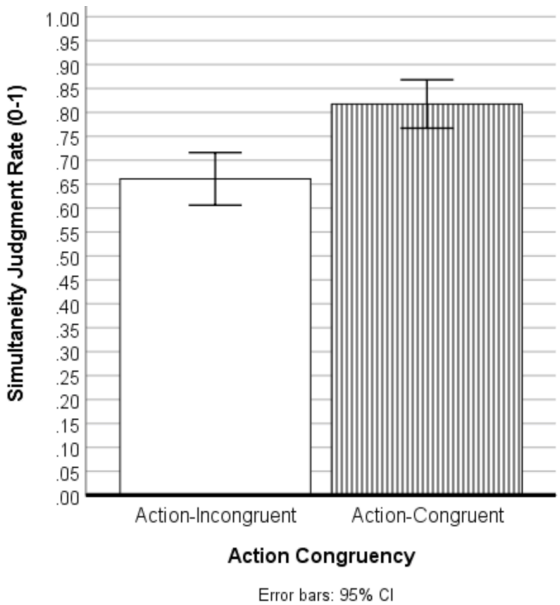

Figure 33. Simultaneity judgment rate by action congruency.

4.1.2 Object Congruency

There was a significant main effect of object congruency on localization biasing, $F(1,27) = 4.405$, $p = .045$, $\eta^2 = .140$. As shown in Figure 34, participants exhibited greater localization biasing when presented with object-congruent stimuli ($M = 2.33$, $SD = 1.80$) compared to object-incongruent stimuli ($M = 2.06$, $SD = 1.82$).

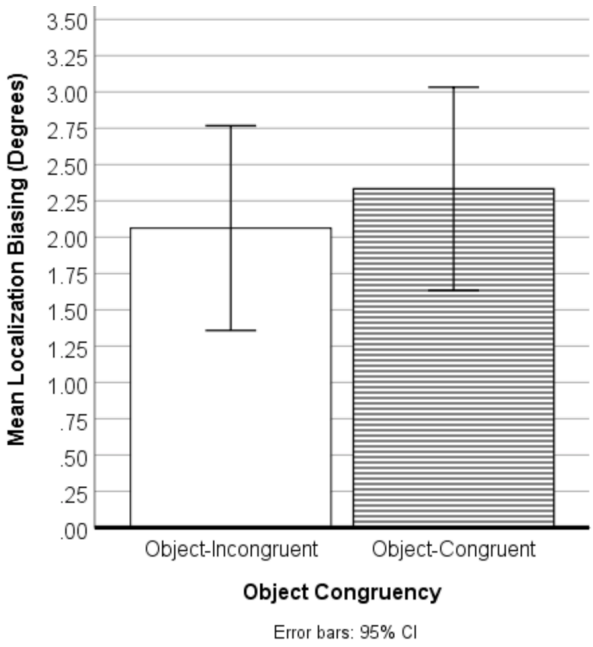

Figure 34. Mean localization biasing by object congruency.

Similarly, there was a significant main effect of action congruency on simultaneity judgment rate, $F(1,27) = 5.558$, $p = .026$, $\eta^2 = .171$. As shown in Figure 35, participants indicated that stimuli were simultaneous at a higher rate in the object-congruent conditions ($M = 0.76$, $SD = 0.13$) compared to the object-incongruent conditions ($M = 0.72$, $SD = 0.12$).

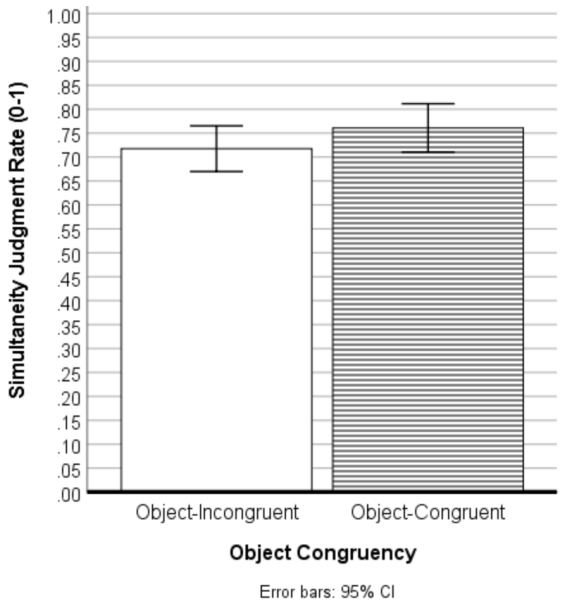

Figure 35. Simultaneity judgment rate by object congruency.

4.2 Hypothesis 2: Comparing Congruency Types

Of the two types of crossmodal congruency, action congruency had the larger impact (see Figure 30 and Figure 31). The planned paired t-test comparing the two partially congruent conditions indicated that localization biasing scores were significantly different in the Action-Congruent-Object-Incongruent condition compared to the Action-Incongruent-Object-Congruent condition, $t(27) = 3.232$, $p = .006$, $d = 0.34$. Localization biasing was greater in the Action-Congruent-Object-Incongruent condition ($M = 2.51$, $SD = 1.82$) compared to the Action-Incongruent-Object-Congruent condition ($M = 1.89$, $SD = 1.88$).

These two conditions also performed differently in terms of simultaneity judgment rates, $t(27) = 4.845$, $p < .001$, $d = 0.83$. Simultaneity judgment rates were higher in the Action-Congruent-Object-Incongruent condition ($M = 0.80$, $SD = 0.13$) compared to the Action-Incongruent-Object-Congruent condition ($M = 0.69$, $SD = 0.14$).

4.3 Hypothesis 3: Congruency Type Interactions and Superadditivity

The interaction between action congruency and object congruency was not significant, $F(1,27) = .007$, $p = .936$, $\eta^2 = .000$. As shown in Figure 36, the planned t-test against the superadditivity threshold indicated that the sum of the two difference scores ($M = 1.18$, $SD = 1.41$) was not significantly different from the fully congruent difference score ($M = 1.16$, $SD = 1.37$), $t(27) = .081$, $p = .936$, $d = 0.01$. This suggests that the two types of congruency take effect as a simple linear summation.

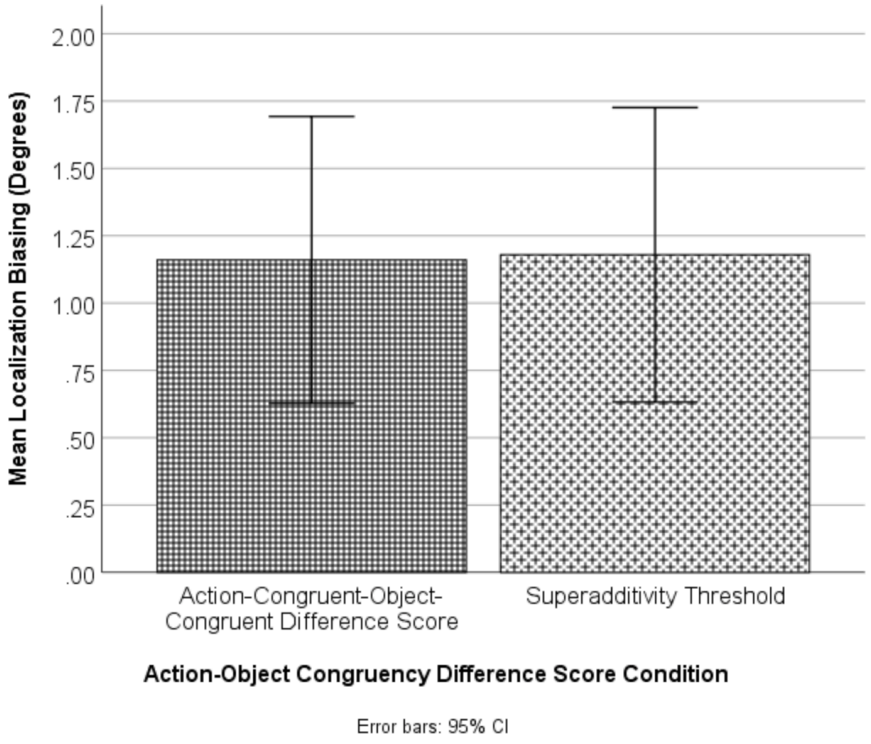

Figure 36. Mean localization biasing by action-object congruency difference score condition.

CHAPTER 5. DISCUSSION

5.1 Implications to Theory

Results indicate that both action and object congruency are valid constructs that have an impact on MSI. The impact of action congruency was larger than the impact of object congruency, via both dependent variables. Setting these effect sizes in the context of other known crossmodal congruency effects, the effect of action congruency was relatively large and the effect of object congruency relatively small. Parise & Spence (2008) observed an effect size of parametric frequency-size congruency on a TOJ task of $\eta^2 = .432$. This was smaller than the effect observed in the present study for action congruency ($\eta^2 = .567$) and larger than the effect of object congruency ($\eta^2 = .171$). Similarly, Parise and Spence (2009) found an effect size of d = 0.36 for frequency: size congruency on spatial ventriloquism, which is smaller than the $\eta^2 = .453$ observation of the effect of action congruency on localization biasing, and larger than the $\eta^2 = .140$ observation of the effect of object congruency on localization biasing (η^2 was not available for direct comparison). These comparisons suggest that, although action congruency is more impactful, both action and object congruency are of similar importance to human perception as parametric congruency effects.

The pattern of results also supports the linear summation model of MSI, and is in accordance with Shams & Kim (2010), Trommershauser, Kording, & Landy (2011). The absence of interaction effects, superadditivity, or subadditivity suggests that some or all action-object congruency effects do not interact, and thus may be investigated and understood, tractably, as independent effects.

5.2 Implications to Practice

Since action and object congruency are both impactful and appear to act independently, both may be individually applied to facilitate MSI in XR systems. Depending on the design context, it may be feasible to adhere to one or the other, but not both. These results indicate that making a practical choice in such situations is acceptable, since the effect of one is not contingent on the other. Taken in the context of previous work supporting the linear summation model, these results support the broader case that "partial" congruency, accumulated through adherence to some congruency types, should be appreciably better for MSI than no congruency at all, and that designers ought to consider adherence to as many known feature-congruency effects as possible. This is not to say that there are not cases of interactions or cue conflict (e.g., Melara and Marks, 1990), but overall an advisable approach may be for designers of XR objects to consider adherence to the variety of congruency effects individually. If it is only feasible to adhere to one type of congruency, action congruency can be expected to be more impactful than object congruency.

However, the task of designing XR objects to be crossmodally congruent could become inflexible if viewed purely through the lens of semantic congruency, and cumbersome or contradictory if viewed purely through the lens of parametric congruency. Semantic effects may not be relevant if the object is novel. Parametric effects do not adhere to a unifying framework and are, as such, difficult to put into practice. For example, it is unclear whether a small but low-elevation object would be congruent with a higher or lower frequency sound.

The framework of action-object congruency, by contrast, provides a non-contradictory way of understanding the type of sounds and visuals that will be crossmodally congruent, and generating adherent stimuli. If, for example, a certain user interface element outside of the user's central vision needed to produce a multisensory alert signal, designers could utilize action-object congruency in the following way.

First, utilize visual cues to create apparent *object* features for the target virtual object (the acted-upon object). These could include roughness and rigidity, the object cues utilized in the present study.

Next, depict a legible *action* that would produce a sound. Although in many cases the depicted action may be inherent to what is being portrayed, in other cases (such as the presently-discussed domain of user interface elements), actions could be created solely for the purpose of supporting action-object congruency. One possibility for supporting action depiction is the use of a universal acting-object with consistent object properties as a physical "cursor," that could strike, rub or scrape target objects. Alternatively, a specific acting-object could be maintained or created for each target object. In either case, actions should depicted as clearly as possible.

Finally, either select from a large number of sound samples, or utilize a procedural sound synthesizer to generate an action-object congruent sound, and then play this sound at the same time and apparent location in space as the depicted event.

Facilitation of MSI is likely to be larger when the action that is depicted is animated in a way that is either physics-inspired or physics-driven. As such, one path toward congruency may be to simulate the visual depiction of the sound-producing event

as well as the produced sound. For example, instead of acting and acted-upon objects being represented by a pre-scripted animation and the playing of a pre-selected sound file, simulated gravity could induce the acting object to fall and impact the acted-upon object in a manner appropriate to its object properties, and a procedural sound synthesizer of the sort designed by Conan et al. (2014) could generate and play an action-object congruent sound in real-time.

In this model, it becomes the task of the designer to design the initial conditions of the simulation in order to produce a desired sound, but the specifics of the sound generation are automated. For example, a designer may decide that an incoming message should sound like two rigid objects impacting at high velocity. They could define the simulated rigidity of the two objects, and then launch the acting-object at high velocity. However, actual sound itself may be slightly different each time it is produced, because the details of the physically simulated interaction may change depending on stochastic elements or differences in the VE. In this model, the XR system designer does not directly design the sound, but instead sets the conditions that will lead to the desired visual and auditory event taking place in an action-object congruent manner.

5.3 Conclusion

This research has described and demonstrated the existence of a new type of crossmodal congruency. Action-object congruency constitutes a widely applicable way of approaching crossmodal congruency that can be used to better understand how humans conduct MSI, and to increase the usability of XR systems.

APPENDIX A. Stimuli Validation Questions

Please rate the extent to which this sound matches the video clip.

- ○ Perfect match
- ○ Very good match
- ○ Good match
- ○ Somewhat good match
- ○ Somewhat poor match
- ○ Poor match
- ○ Very poor match
- ○ Not at all matched

Please indicate the extent to which you agree or disagree with the following statement: The objects in the video remind me of other objects that I have seen, outside of this study.

- ○ Strongly agree
- ○ Agree
- ○ Somewhat agree
- ○ Neither agree nor disagree
- ○ Somewhat disagree
- ○ Disagree
- ○ Strongly disagree

Please indicate the extent to which you agree or disagree with the following statement: It would be easy to name the material that the objects in the video are made of.

- ○ Strongly agree
- ○ Agree
- ○ Somewhat agree
- ○ Neither agree nor disagree
- ○ Somewhat disagree
- ○ Disagree
- ○ Strongly disagree

If the objects did remind you of anything, what was it?

_____

APPENDIX B. Simulator Sickness Questionnaire

Please indicate the extent to which you are experiencing the following symptoms:

	None	Slight	Moderate	Severe
General discomfort	○	○	○	○
Fatigue	○	○	○	○
Headache	○	○	○	○
Eye strain	○	○	○	○
Difficulty focusing	○	○	○	○
Increased salivation	○	○	○	○
Sweating	○	○	○	○
Nausea	○	○	○	○
Difficulty concentrating	○	○	○	○
"Fullness" of the head	○	○	○	○
Blurred vision	○	○	○	○
Dizziness (eyes open)	○	○	○	○
Dizziness (eyes closed)	○	○	○	○
Vertigo	○	○	○	○
Stomach awareness	○	○	○	○
Burping	○	○	○	○

APPENDIX C. Goldsmiths Musical Sophistication Index

	Completely Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Completely Agree
I spend a lot of my free time doing music-related activities.	○	○	○	○	○	○	○
I enjoy writing about music, for example on blogs and forums.	○	○	○	○	○	○	○
If somebody starts singing a song I don't know, I can usually join in.	○	○	○	○	○	○	○
I can sing or play music from memory.	○	○	○	○	○	○	○
I am able to hit the right notes when I sing along with a recording.	○	○	○	○	○	○	○
I can compare and discuss differences between two performances or versions of the same piece of music.	○	○	○	○	○	○	○
I have never been complimented for my talents as a musical performer.	○	○	○	○	○	○	○

	Completely Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Completely Agree
I often read or search the internet for things related to music.	○	○	○	○	○	○	○
I am not able to sing in harmony when somebody is singing a familiar tune.	○	○	○	○	○	○	○
I am able to identify what is special about a given musical piece.	○	○	○	○	○	○	○
When I sing, I have no idea whether I'm in tune or not.	○	○	○	○	○	○	○
Music is kind of an addiction for me - I couldn't live without it.	○	○	○	○	○	○	○
I don’t like singing in public because I’m afraid that I would sing wrong notes.	○	○	○	○	○	○	○
I would not consider myself a musician.	○	○	○	○	○	○	○
After hearing a new song two or three times, I can usually sing it by myself.	○	○	○	○	○	○	○

APPENDIX D. Demographic and VR Questions

How often did the two cubes appear to interact within your field of view?

- ○ Every time
- ○ Most of the time
- ○ About half of the time
- ○ Some of the time
- ○ Never

When looking straight ahead within the headset, were the visuals clear?

- ○ Definitely yes
- ○ Probably yes
- ○ Might or might not
- ○ Probably not
- ○ Definitely not

Were you able to hear a sound during every trial?

- ○ Definitely yes
- ○ Probably yes
- ○ Might or might not
- ○ Probably not
- ○ Definitely not

What is your age?

_____

What is your gender?

- ○ Male
- ○ Female
- ○ Other (Please Specify)

_____

Do you have any hearing impairments? In particular, do you have selective hearing loss in one ear?

- ○ No
- ○ Don't know/ prefer not to answer
- ○ Yes (Please specify) _____

Do you own a virtual reality device?

- ○ Yes, and I use it regularly
- ○ Yes, and I use it occasionally
- ○ Yes, but I do not use it
- ○ No

How much total time have you spent using virtual reality systems?

- ○ More than 24 hours
- ○ 5-24 hours
- ○ 1-5 hours
- ○ 0-1 hours
- ○ None at all

REFERENCES

Aslin, R. N., & Newport, E. L. (2008). What statistical learning can and can't tell us about language acquisition. In J. Colombo, P. McCardle, & L. Freund (Eds.). *Infant pathways to language: Methods, models, and research directions*. Mahwah, NJ: Erlbaum.

Bach, D. R., & Dolan, R. J. (2012). Knowing how much you don't know: a neural organization of uncertainty estimates. *Nature Reviews Neuroscience*, *13*(8), 572.

Baddeley, A., Allen, R. J., & Hitch, G. (2010). Investigating the episodic buffer. *Psychologica Belgica*, *50*(3), 223–243.

Bailey, H. D., Mullaney, A. B., Gibney, K. D., & Kwakye, L. D. (2018). Audiovisual Integration Varies With Target and Environment Richness in Immersive Virtual Reality. *Multisensory Research*, *31*(7), 689–713.

Beauregard Cazabon, D. (2016). The role of amplitude envelope in audio-visual perception: testing the effect of amplitude envelope in spatial ventriloquism. (Doctoral dissertation – McMaster University).

Begault, D. R., Wenzel, E. M., & Anderson, M. R. (2001). Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. *Journal of the Audio Engineering Society*, *49*(10), 904–916.

Belardinelli, M. O., Sestieri, C., Di Matteo, R., Delogu, F., Del Gratta, C., Ferretti, A., … Romani, G. L. (2004). Audio-visual crossmodal interactions in environmental perception: an fMRI investigation. *Cognitive Processing*, *5*(3), 167–174.

Bernstein, I. H., & Edelstein, B. A. (1971). Effects of some variations in auditory input upon visual choice reaction time. *Journal of Experimental Psychology*, *87*(2), 241.

Bertelson, P., & Aschersleben, G. (1998). Automatic visual bias of perceived auditory location. *Psychonomic Bulletin & Review*, *5*(3), 482–489.

Bhat, J., Miller, L. M., Pitt, M. A., & Shahin, A. J. (2014). Putative mechanisms mediating tolerance for audiovisual stimulus onset asynchrony. *Journal of Neurophysiology*, *113*(5), 1437–1450.

Bidelman, G. M. (2016). Musicians have enhanced audiovisual multisensory binding: experience-dependent effects in the double-flash illusion. *Experimental Brain Research*, *234*(10), 3037–3047.

Bien, N., ten Oever, S., Goebel, R., & Sack, A. T. (2012). The sound of size: crossmodal binding in pitch-size synesthesia: a combined TMS, EEG and psychophysics study. *NeuroImage*, *59*(1), 663–672.

Bizley, J. K., Jones, G. P., & Town, S. M. (2016). Where are multisensory signals combined for perceptual decision-making? *Current Opinion in Neurobiology*, *40*, 31–37.

Bizley, J. K., Maddox, R. K., & Lee, A. K. C. (2016). Defining auditory-visual objects: Behavioral tests and physiological mechanisms. *Trends in Neurosciences*, *39*(2), 74–85.

Blauert, J. (1997). *Spatial hearing: the psychophysics of human sound localization*. Cambridge, MA: MIT press.

Bonetti, L., & Costa, M. (2018). Pitch-verticality and pitch-size cross-modal interactions. *Psychology of Music*, *46*(3), 340–356.

Boyle, S. C., Kayser, C., & Ince, R. A. A. (2018). Early Neural Correlates of an Auditory Pitch-Visual Size Cross-modal Association. *BioRxiv*, 423939.

Bronkhorst, A. W. (1995). Localization of real and virtual sound sources. *The Journal of the Acoustical Society of America*, *98*(5), 2542–2553.

Bronkhorst, A. W., Veltman, J. A., & Van Breda, L. (1996). Application of a three-dimensional auditory display in a flight task. *Human Factors*, *38*(1), 23–33.

Brunetti, R., Indraccolo, A., Del Gatto, C., Spence, C., & Santangelo, V. (2018). Are crossmodal correspondences relative or absolute? Sequential effects on speeded classification. *Attention, Perception, & Psychophysics*, *80*(2), 527–534.

Bruns, P., & Röder, B. (2019). Spatial and frequency specificity of the ventriloquism aftereffect revisited. *Psychological Research*, *83*(7), 1400–1415.

Burr, D., Silva, O., Cicchini, G. M., Banks, M. S., & Morrone, M. C. (2009). Temporal mechanisms of multimodal binding. *Proceedings of the Royal Society B: Biological Sciences*, *276*(1663), 1761–1769.

Cabe, P. A., Bochtler, K. S., & Neuhoff, J. G. (2018). Squeaky wheels: Friction-generated sound supports auditory differentiation and scaling of rotating ellipse shapes. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(7), 1054.

Carello, C., Wagman, J. B., & Turvey, M. T. (2005). Acoustic specification of object properties. *Moving Image Theory: Ecological Considerations*, 79–104.

Carnevale, Michael J, & Harris, L. R. (2016). Which direction is up for a high pitch? *Multisensory Research*, *29*(1–3), 113–132.

Carnevale, Michael James. (2015). What's Up with High-and Low-Pitched Sounds? Reference Frames used in the Crossmodal Correspondence Between Auditory Pitch and Visuospatical Height. (Doctoral disserrtation – York University).

Cellini, C., Kaim, L., & Drewing, K. (2013). Visual and haptic integration in the estimation of softness of deformable objects. *I-Perception*, *4*(8), 516–531.

Chalk, M., Seitz, A. R., & Seriès, P. (2010). Rapidly learned stimulus expectations alter perception of motion. *Journal of Vision*, *10*(8), 2.

Chan, Z. P. Y., & Dyson, B. J. (2015). The effects of association strength and cross-modal correspondence on the development of multimodal stimuli. *Attention, Perception, & Psychophysics*, *77*(2), 560–570.

Chen, Y.-C., & Spence, C. (2017). Assessing the role of the 'unity assumption'on multisensory integration: A review. *Frontiers in Psychology*, *8*, 445.

Chiou, R., Stelter, M., & Rich, A. N. (2013). Beyond colour perception: auditory--visual synaesthesia induces experiences of geometric objects in specific locations. *Cortex*, *49*(6), 1750–1763.

Chuen, L., & Schutz, M. (2016). The unity assumption facilitates cross-modal binding of musical, non-speech stimuli: The role of spectral and amplitude envelope cues. *Attention, Perception, & Psychophysics*, *78*(5), 1512–1528.

Clavagnier, S., Falchier, A., & Kennedy, H. (2004). Long-distance feedback projections to area V1: implications for multisensory integration, spatial awareness, and visual consciousness. *Cognitive, Affective, & Behavioral Neuroscience*, *4*(2), 117–126.

Conan, S., Aramaki, M., Kronland-Martinet, R., Thoret, E., & Ystad, S. (2012). Perceptual differences between sounds produced by different continuous interactions. In *Proceedings of the Acoustics 2012 Nantes Conference*, (pp. 581-586).

Conan, S., Derrien, O., Aramaki, M., Ystad, S., & Kronland-Martinet, R. (2014). A synthesis model with intuitive control capabilities for rolling sounds. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(8), 1260–1273.

Conan, S., Thoret, E., Aramaki, M., Derrien, O., Gondre, C., Kronland-Martinet, R., & Ystad, S. (2013). Navigating in a space of synthesized interaction-sounds: Rubbing, scratching and rolling sounds. In *16th International Conference on Digital Audio Effects (DAFx),* (pp. 202–209).

Connolly, K. (2014). Multisensory perception as an associative learning process. *Frontiers in Psychology*, *5*, 1095.

Coward, S. W., & Stevens, C. J. (2004). Extracting meaning from sound: Nomic mappings, everyday listening, and perceiving object size from frequency. *The Psychological Record*, *54*(3), 349–364.

Darvishi, A., Munteanu, E., Guggiana, V., Schauer, H., Motavalli, M., & Rauterberg, M.

(1995). Designing environmental sounds based on the results of interaction between objects in the real world. In *Human—Computer Interaction* (pp. 38–42). Springer.

Deroy, O., & Spence, C. (2013). Are we all born synaesthetic? Examining the neonatal synaesthesia hypothesis. *Neuroscience & Biobehavioral Reviews*, *37*(7), 1240–1253.

Doehrmann, O., & Naumer, M. J. (2008). Semantics and the multisensory brain: how meaning modulates processes of audio-visual integration. *Brain Research*, *1242*, 136–150.

Ernst, M. O. (2007). Learning to integrate arbitrary signals from vision and touch. *Journal of Vision*, *7*(5), 7.

Ferris, T. K., & Sarter, N. B. (2008). Cross-modal links among vision, audition, and touch in complex environments. *Human Factors*, *50*(1), 17–26.

Fifer, J. M., Barutchu, A., Shivdasani, M. N., & Crewther, S. G. (2013). Verbal and novel multisensory associative learning in adults. *F1000Research*, *2*.

Froyen, D., Van Atteveldt, N., Bonte, M., & Blomert, L. (2008). Cross-modal enhancement of the MMN to speech-sounds indicates early and automatic integration of letters and speech-sounds. *Neuroscience Letters*, *430*(1), 23–28.

Gallace, A., & Spence, C. (2006). Multisensory synesthetic interactions in the speeded classification of visual size. *Perception & Psychophysics*, *68*(7), 1191–1203.

Gao, Z., Wu, F., Qiu, F., He, K., Yang, Y., & Shen, M. (2017). Bindings in working memory: The role of object-based attention. *Attention, Perception, & Psychophysics*, *79*(2), 533–552.

Gaver, W. W. (1993). What in the world do we hear?: An ecological approach to auditory event perception. *Ecological Psychology*, *5*(1), 1–29.

Gerhard, H. E., Wichmann, F. A., & Bethge, M. (2013). How sensitive is the human visual system to the local statistics of natural images? *PLoS Computational Biology*, *9*(1), e1002873.

Gilbert, C. D., Sigman, M., & Crist, R. E. (2001). The neural basis of perceptual learning. *Neuron*, *31*(5), 681–697.

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, *14*(7), 926.

Glicksohn, A., & Cohen, A. (2013). The role of cross-modal associations in statistical learning. *Psychonomic Bulletin & Review*, *20*(6), 1161–1169.

Graham, R. E. (2017). Music to Our Eyes: Assessing the Role of Experience for

Multisensory Integration in Music Perception. (Doctoral dissertation – Southern Illinous University Carbondale).

Grassi, M. (2005). Do we hear size or sound? Balls dropped on plates. *Perception & Psychophysics*, *67*(2), 274–284.

Grassi, M., & Casco, C. (2010). Audiovisual bounce-inducing effect: When sound congruence affects grouping in vision. *Attention, Perception, & Psychophysics*, *72*(2), 378–386.

Grassi, M., Pastore, M., & Lemaitre, G. (2013). Looking at the world with your ears: How do we get the size of an object from its sound? *Acta Psychologica*, *143*(1), 96–104.

Guzman-Martinez, E., Ortega, L., Grabowecky, M., Mossbridge, J., & Suzuki, S. (2012). Interactive coding of visual spatial frequency and auditory amplitude-modulation rate. *Current Biology*, *22*(5), 383–388.

Habets, B., Bruns, P., & Röder, B. (2017). Experience with crossmodal statistics reduces the sensitivity for audio-visual temporal asynchrony. *Scientific Reports*, *7*(1), 1486.

Hamilton-Fletcher, G., Pisanski, K., Reby, D., Stefańczyk Michałand Ward, J., & Sorokowska, A. (2018). The role of visual experience in the emergence of cross-modal correspondences. *Cognition*, *175*, 114–121.

Hartmann, C., Lazar, A., & Triesch, J. (2014). Where's the noise? Key features of neuronal variability and inference emerge from self-organized learning. *BioRxiv*, 11296.

Hein, G., Doehrmann, O., Müller, N. G., Kaiser, J., Muckli, L., & Naumer, M. J. (2007). Object familiarity and semantic congruency modulate responses in cortical audiovisual integration areas. *Journal of Neuroscience*, *27*(30), 7881–7887.

Hirsh, I. J., & Sherrick Jr, C. E. (1961). Perceived order in different sense modalities. *Journal of Experimental Psychology*, *62*(5), 423.

Honbolygó, F., Veller, L., & Csépe, V. (2012). Ventriloquism aftereffect in a virtual audio-visual environment. In *2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)* (pp. 475–478).

Houben, M. M. J., Kohlrausch, A., & Hermes, D. J. (2004). Perception of the size and speed of rolling balls by sound. *Speech Communication*, *43*(4), 331–345.

Hubbard, T. L. (1996). Synesthesia-like mappings of lightness, pitch, and melodic interval. *The American Journal of Psychology*, 219–238.

Hummel, J. E. (2000). Where view-based theories break down: The role of structure in shape perception and object recognition. *Cognitive Dynamics: Conceptual Change in Humans and Machines*, 157–185.

Iwaki, M., & Chigira, Y. (2016). Compensation of sound source direction perceived through consumer-grade bone-conduction headphones by modifying ILD and ITD. In *2016 IEEE 5th Global Conference on Consumer Electronics* (pp. 1–4).

J Steenson, C., Rodger, M., & Matthew, W. (2015). Bringing sounds into use: thinking of sounds as materials and a sketch of auditory affordances. *The Open Psychology Journal*, *8*(1).

Jonas, C., Spiller, M. J., & Hibbard, P. (2017). Summation of visual attributes in auditory--visual crossmodal correspondences. *Psychonomic Bulletin & Review*, *24*(4), 1104–1112.

Jordan, K. E., Clark, K., & Mitroff, S. R. (2010). See an object, hear an object file: Object correspondence transcends sensory modality. *Visual Cognition*, *18*(4), 492–503.

Karwoski, T. F., Odbert, H. S., & Osgood, C. E. (1942). Studies in synesthetic thinking: II. The role of form in visual responses to music. *The Journal of General Psychology*, *26*(2), 199–222.

Keetels, M., & Vroomen, J. (2011). No effect of synesthetic congruency on temporal ventriloquism. *Attention, Perception, & Psychophysics*, *73*(1), 209–218.

Keil, J., & Senkowski, D. (2018). Neural oscillations orchestrate multisensory processing. *The Neuroscientist*, *24*(6), 609–626.

Kennedy, R. S., Lane, N. E., Berbaum, K. S., & Lilienthal, M. G. (1993). Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness. *The International Journal of Aviation Psychology*, *3*(3), 203–220.

Kim, R., Seitz, A., Feenstra, H., & Shams, L. (2009). Testing assumptions of statistical learning: is it long-term and implicit? *Neuroscience Letters*, *461*(2), 145–149.

Knapp, J. M., & Loomis, J. M. (2004). Limited field of view of head-mounted displays is not the cause of distance underestimation in virtual environments. *Presence: Teleoperators & Virtual Environments*, *13*(5), 572–577.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, *27*(12), 712–719.

Koelewijn, T., Bronkhorst, A., & Theeuwes, J. (2010). Attention and the multiple stages of multisensory integration: A review of audiovisual studies. *Acta Psychologica*, *134*(3), 372–384.

Koppen, C., Alsius, A., & Spence, C. (2008). Semantic congruency and the Colavita visual dominance effect. *Experimental Brain Research*, *184*(4), 533–546.

Kruijff, E., Swan, J. E., & Feiner, S. (2010). Perceptual issues in augmented reality revisited. In *2010 IEEE International Symposium on Mixed and Augmented Reality*

(pp. 3–12).

Kytö, M., Kusumoto, K., & Oittinen, P. (2015). The ventriloquist effect in augmented reality. In *2015 IEEE International Symposium on Mixed and Augmented Reality* (pp. 49–53).

Lakatos, S., McAdams, S., & Caussé, R. (1997). The representation of auditory source characteristics: Simple geometric form. *Perception & Psychophysics*, *59*(8), 1180–1190.

Laurienti, P. J., Kraft, R. A., Maldjian, J. A., Burdette, J. H., & Wallace, M. T. (2004). Semantic congruence is a critical factor in multisensory behavioral performance. *Experimental Brain Research*, *158*(4), 405–414.

Laurienti, P. J., Wallace, M. T., Maldjian, J. A., Susi, C. M., Stein, B. E., & Burdette, J. H. (2003). Cross-modal sensory processing in the anterior cingulate and medial prefrontal cortices. *Human Brain Mapping*, *19*(4), 213–223.

Lee, W., & Sato, M. (2001). Visual perception of texture of textiles. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de La Couleur*, *26*(6), 469–477.

Lemaitre, G., & Heller, L. M. (2012). Auditory perception of material is fragile while action is strikingly robust. *The Journal of the Acoustical Society of America*, *131*(2), 1337–1348.

Loomis, J. M., Lippa, Y., Klatzky, R. L., & Golledge, R. G. (2002). Spatial updating of locations specified by 3-D sound and spatial language. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(2), 335.

Luckett, E. (2018). *A Quantitative Evaluation of the HTC Vive for Virtual Reality Research*. (Doctoral dissertation, The University of Mississippi).

Ludwig, V. U., Adachi, I., & Matsuzawa, T. (2011). Visuoauditory mappings between high luminance and high pitch are shared by chimpanzees (Pan troglodytes) and humans. *Proceedings of the National Academy of Sciences*, *108*(51), 20661–20665.

Lutfi, R. A. (2001). Auditory detection of hollowness. *The Journal of the Acoustical Society of America*, *110*(2), 1010–1019.

Macaluso, E., Noppeney, U., Talsma, D., Vercillo, T., Hartcher-O'Brien, J., & Adam, R. (2016). The curious incident of attention in multisensory integration: bottom-up vs. top-down. *Multisensory Research*, *29*(6–7), 557–583.

Makovac, E., & Gerbino, W. (2010). Sound-shape congruency affects the multisensory response enhancement. *Visual Cognition, 18*, 133–137.

Marks, L. E. (1974). On associations of light and sound: The mediation of brightness, pitch, and loudness. *The American Journal of Psychology*, 173–188.

Martino, G., & Marks, L. E. (1999). Perceptual and linguistic interactions in speeded classification: Tests of the semantic coding hypothesis. *Perception*, *28*(7), 903–923.

McGovern, D. P., Roudaia, E., Newell, F. N., & Roach, N. W. (2016). Perceptual learning shapes multisensory causal inference via two distinct mechanisms. *Scientific Reports*, *6*, 24673.

Melara, R. D. (1989). Similarity relations among synesthetic stimuli and their attributes. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(2), 212.

Melara, R. D., & Marks, L. E. (1990). HARD and SOFT interacting dimensions: Differential effects of dual context on classification. *Perception & Psychophysics*, *47*(4), 307–325.

Melara, R. D., & O'Brien, T. P. (1987). Interaction between synesthetically corresponding dimensions. *Journal of Experimental Psychology: General*, *116*(4), 323.

Mitchel, A. D., Christiansen, M. H., & Weiss, D. J. (2014). Multimodal integration in statistical learning: evidence from the McGurk illusion. *Frontiers in Psychology*, *5*, 407.

Molholm, S., Ritter, W., Javitt, D. C., & Foxe, J. J. (2004). Multisensory visual--auditory object recognition in humans: a high-density electrical mapping study. *Cerebral Cortex*, *14*(4), 452–465.

Monache, S. D., Polotti, P., & Rocchesso, D. (2010). A toolkit for explorations in sonic interaction design. In *Proceedings of the 5th audio mostly conference: a conference on interaction with sound*, (pp. 1-7).

Morein-Zamir, S., Soto-Faraco, S., & Kingstone, A. (2003). Auditory capture of vision: examining temporal ventriloquism. *Cognitive Brain Research*, *17*(1), 154–163.

Mudd, S. A. (1963). Spatial stereotypes of four dimensions of pure tone. *Journal of Experimental Psychology*, *66*(4), 347.

Mullan, E. (2009). Driving sound synthesis from a physics engine. In *2009 International IEEE Consumer Electronics Society's Games Innovations Conference*, 1–9.

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). Measuring the facets of musicality: The Goldsmiths Musical Sophistication Index (Gold-MSI). *Personality and Individual Differences*, *60*, S35.

Munoz, N. E., & Blumstein, D. T. (2012). Multisensory perception in uncertain environments. *Behavioral Ecology*, *23*(3), 457–462.

Nastase, S. A., Davis, B., & Hasson, U. (2018). Cross-modal and non-monotonic representations of statistical regularity are encoded in local neural response patterns. *NeuroImage*, *173*, 509–517.

Nelson, W. T., Hettinger, L. J., Cunningham, J. A., Brickman, B. J., Haas, M. W., & McKinley, R. L. (1998). Effects of localized auditory information on visual target detection performance using a helmet-mounted display. *Human Factors*, *40*(3), 452–460.

Nidiffer, A. R., Diederich, A., Ramachandran, R., & Wallace, M. T. (2018). Multisensory perception reflects individual differences in processing temporal correlations. *Scientific Reports*, *8*.

O'leary, A., & Rhodes, G. (1984). Cross-modal effects on visual and auditory object perception. *Perception & Psychophysics*, *35*(6), 565–569.

Olson, I. R., Gatenby, J. C., & Gore, J. C. (2002). A comparison of bound and unbound audio--visual information processing in the human cerebral cortex. *Cognitive Brain Research*, *14*(1), 129–138.

Paraskevopoulos, E., Kuchenbuch, A., Herholz, S. C., & Pantev, C. (2012). Evidence for training-induced plasticity in multisensory brain structures: an MEG study. *PloS One*, *7*(5), e36534.

Parise, C., & Spence, C. (2008). Synesthetic congruency modulates the temporal ventriloquism effect. *Neuroscience Letters*, *442*(3), 257–261.

Parise, C. V. (2012). Signal compatibility as a modulatory factor for audiovisual multisensory integration. *integration*, *22*, 46-49.

Parise, C. V., & Spence, C. (2009). 'When birds of a feather flock together': Synesthetic correspondences modulate audiovisual integration in non-synesthetes. *PLoS One*, *4*(5), e5664.

Parise, C. V, & Ernst, M. O. (2017). Noise, multisensory integration, and previous response in perceptual disambiguation. *PLoS Computational Biology*, *13*(7), e1005546.

Parise, C. V, Knorre, K., & Ernst, M. O. (2014). Natural auditory scene statistics shapes human spatial hearing. *Proceedings of the National Academy of Sciences*, *111*(16), 6104–6108.

Peters, M. A. K., Balzer, J., & Shams, L. (2015). Smaller= denser, and the brain knows it: Natural statistics of object density shape weight expectations. *PloS One*, *10*(3), e0119794.

Piazza, E. A., Denison, R. N., & Silver, M. A. (2018). Recent cross-modal statistical learning influences visual perceptual selection. *Journal of Vision*, *18*(3), 1.

Piemo, A. C., Caria, A., & Castiello, U. (2006). Crossmodal binding in localizing objects outside the field of view. *Visual Cognition*, *13*(2), 223–246.

Pisanski, K., Isenstein, S. G. E., Montano, K. J., O'Connor, J. J. M., & Feinberg, D. R. (2017). Low is large: spatial location and pitch interact in voice-based body size estimation. *Attention, Perception, & Psychophysics*, *79*(4), 1239–1251.

Pitkow, X., & Angelaki, D. E. (2017). Inference in the brain: statistics flowing in redundant population codes. *Neuron*, *94*(5), 943–953.

Pourtois, G., & de Gelder, B. (2002). Semantic factors influence multisensory pairing: a transcranial magnetic stimulation study. *Neuroreport*, *13*(12), 1567–1573.

Powell, D., Merrick, M. A., Lu, H., & Holyoak, K. J. (2016). Causal competition based on generic priors. *Cognitive Psychology*, *86*, 62–86.

Preis, A., & Klawiter, A. (2005). The audition of natural sounds--its levels and relevant experiments. In *Proceedings of Forum Acusticum, Kraków*, 1595–1599.

Pruvost, L., Scherrer, B., Aramaki, M., Ystad, S., & Kronland-Martinet, R. (2015). Perception-based interactive sound synthesis of morphing solids' interactions. In *SIGGRAPH Asia 2015 Technical Briefs* (p. 17).

Putzar, L., Goerendt, I., Lange, K., Rösler, F., & Röder, B. (2007). Early visual deprivation impairs multisensory interactions in humans. *Nature Neuroscience*, *10*(10), 1243.

Raghuvanshi, N., Narain, R., & Lin, M. C. (2009). Efficient and accurate sound propagation using adaptive rectangular decomposition. *IEEE Transactions on Visualization and Computer Graphics*, *15*(5), 789–801.

Rahnev, D. (2017a). The case against full probability distributions in perceptual decision making. *BioRxiv*, 108944.

Rahnev, D. (2017b). Top-down control of perceptual decision making by the prefrontal cortex. *Current Directions in Psychological Science*, *26*(5), 464–469.

Rahnev, D., Lau, H., & de Lange, F. P. (2011). Prior expectation modulates the interaction between sensory and prefrontal regions in the human brain. *Journal of Neuroscience*, *31*(29), 10741–10748.

Rinaldi, L., Lega, C., Cattaneo, Z., Girelli, L., & Bernardi, N. F. (2016). Grasping the sound: Auditory pitch influences size processing in motor planning. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(1), 11.

Rocchesso, D. (2004). Physically-based sounding objects, as we develop them today. *Journal of New Music Research*, *33*(3), 305–313.

Roffler, S. K., & Butler, R. A. (1968). Localization of tonal stimuli in the vertical plane.

The Journal of the Acoustical Society of America, *43*(6), 1260–1266.

Salter, T. G., Sugden, B., Deptford, D., Crocco, R., Keane, B., Massey, L., … others. (2016). Indicating out-of-view augmented reality images. Google Patents.

Sapir, E. (1929). A study in phonetic symbolism. *Journal of Experimental Psychology*, *12*(3), 225.

Savioja, L., Huopaniemi, J., Lokki, T., & Väänänen, R. (1999). Creating interactive virtual acoustic environments. *Journal of the Audio Engineering Society*, *47*(9), 675–705.

Schneider, T. R., Debener, S., Oostenveld, R., & Engel, A. K. (2008). Enhanced EEG gamma-band activity reflects multisensory semantic matching in visual-to-auditory object priming. *Neuroimage*, *42*(3), 1244–1254.

Scott, S. K. (2005). Auditory processing—speech, space and auditory objects. *Current Opinion in Neurobiology*, *15*(2), 197–201.

Seitz, A., & Leclercq, V. (2012). What level of processing is required to form fast-task irrelevant perceptual learning? *Perception ECVP Abstract*, *41*, 172.

Seitz, A. R., Kim, R., van Wassenhove, V., & Shams, L. (2007). Simultaneous and independent acquisition of multisensory and unisensory associations. *Perception*, *36*(10), 1445–1453.

Shams, L., & Kim, R. (2010). Crossmodal influences on visual perception. *Physics of Life Reviews*, *7*(3), 269–284.

Siebold, A. (2009). *The influence of degraded stimuli on audio-visual integration*. University of Twente.

Slutsky, D. A., & Recanzone, G. H. (2001). Temporal and spatial dependency of the ventriloquism effect. *Neuroreport*, *12*(1), 7–10.

Smith, E. L., Grabowecky, M., & Suzuki, S. (2007). Auditory-visual crossmodal integration in perception of face gender. *Current Biology*, *17*(19), 1680–1685.

Sotiropoulos, G., Seitz, A. R., & Seriès, P. (2011). Changing expectations about speed alters perceived motion direction. *Current Biology*, *21*(21), R883--R884.

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, *73*(4), 971–995.

Spence, C., Ngo, M. K., Lee, J.-H., & Tan, H. (2010). Solving the correspondence problem in haptic/multisensory interface design. *Advances in Haptics*, 47–74.

Spence, C., & Parise, C. V. (2012). The cognitive neuroscience of crossmodal correspondences. *I-Perception*, *3*(7), 410–412.

Spence, C., Sanabria, D., & Soto-Faraco, S. (2007). Intersensory Gestalten and crossmodal scene perception. *Psychology of Beauty and Kansei: New Horizons of Gestalt Perception*, 519–579.

Spence, C., & Squire, S. (2003). Multisensory integration: maintaining the perception of synchrony. *Current Biology*, *13*(13), R519--R521.

Stein, B. E., & Stanford, T. R. (2008). Multisensory integration: current issues from the perspective of the single neuron. *Nature Reviews Neuroscience*, *9*(4), 255.

Stevenson, R. A., Wallace, M. T., & Altieri, N. (2014). The interaction between stimulus factors and cognitive factors during multisensory integration of audiovisual speech. *Frontiers in Psychology*, *5*, 352.

Stoelinga, C. N. J. (2007). *A psychomechanical study of rolling sounds*. ENSTA ParisTech.

Stoelinga, C. N. J., & Lutfi, R. A. (2011). Modeling manner of contact in the synthesis of impact sounds for perceptual research. *The Journal of the Acoustical Society of America*, *130*(2), EL62--EL68.

Suied, C., & Viaud-Delmon, I. (2009). Auditory-visual object recognition time suggests specific processing for animal sounds. *PloS One*, *4*(4), e5256.

Tanabe, H. C., Honda, M., & Sadato, N. (2005). Functionally segregated neural substrates for arbitrary audiovisual paired-association learning. *Journal of Neuroscience*, *25*(27), 6409–6418.

Ten Oever, S., Sack, A. T., Wheat, K. L., Bien, N., & Van Atteveldt, N. (2013). Audio-visual onset differences are used to determine syllable identity for ambiguous audio-visual stimulus pairs. *Frontiers in Psychology*, *4*, 331.

Thomas, R. L., Nardini, M., & Mareschal, D. (2017). The impact of semantically congruent and incongruent visual information on auditory object recognition across development. *Journal of Experimental Child Psychology*, *162*, 72–88.

Thoret, E., Aramaki, M., Kronland-Martinet, R., Velay, J.-L., & Ystad, S. (2014). From sound to shape: Auditory perception of drawing movements. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(3), 983.

Tiest, W. M. B., & Kappers, A. M. L. (2007). Haptic and visual perception of roughness. *Acta Psychologica*, *124*(2), 177–189.

Trommershauser, J., Kording, K., & Landy, M. S. (2011). *Sensory cue integration*. Oxford University Press.

Ursino, M., Cuppini, C., & Magosso, E. (2017). Multisensory Bayesian inference depends on synapse maturation during training: theoretical analysis and neural modeling implementation. *Neural Computation*, *29*(3), 735–782.

Van Den Doel, K., Kry, P. G., & Pai, D. K. (2001). FoleyAutomatic: physically-based sound effects for interactive simulation and animation. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques* (pp. 537–544).

Van Der Hoort, B., & Ehrsson, H. H. (2016). Illusions of having small or large invisible bodies influence visual perception of object size. *Scientific Reports*, *6*, 34530.

Van Wanrooij, M. M., Bremen, P., & John Van Opstal, A. (2010). Acquired prior knowledge modulates audiovisual integration. *European Journal of Neuroscience*, *31*(10), 1763–1771.

Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, *45*(3), 598–607.

van Wassenhove, V., and Schroeder, C. E. (2012). Multisensory role of human auditory cortex. In D. Poeppel, T. Overath, A. N. Popper, and R. R. Fay (Eds.). *The Human Auditory Cortex, Springer Handbook of Auditory Research*. New York, NY: Springer, 295–331.

Vatakis, A., Ghazanfar, A. A., & Spence, C. (2008). Facilitation of multisensory integration by the "unity effect" reveals that speech is special. *Journal of Vision*, *8*(9), 14.

Vatakis, A., & Spence, C. (2007). Crossmodal binding: Evaluating the "unity assumption" using audiovisual speech stimuli. *Perception & Psychophysics*, *69*(5), 744–756.

Von Kriegstein, K., & Giraud, A.-L. (2006). Implicit multisensory associations influence voice recognition. *PLoS Biology*, *4*(10), e326.

Vroomen, J., & Keetels, M. (2006). The spatial constraint in intersensory pairing: No role in temporal ventriloquism. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(4), 1063.

Walker, L., & Walker, P. (2016). Cross-sensory mapping of feature values in the size--brightness correspondence can be more relative than absolute. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(1), 138.

Walker, L., Walker, P., & Francis, B. (2012). A common scheme for cross-sensory correspondences across stimulus domains. *Perception*, *41*(10), 1186–1192.

Wallace, M. T., Perrault, T. J., Hairston, W. D., & Stein, B. E. (2004). Visual experience is necessary for the development of multisensory integration. *Journal of Neuroscience*, *24*(43), 9580–9584.

Wallace, M. T., Roberson, G. E., Hairston, W. D., Stein, B. E., Vaughan, J. W., & Schirillo, J. A. (2004). Unifying multisensory signals across time and space. *Experimental Brain Research*, *158*(2), 252–258.

Wallace, M. T., & Stein, B. E. (2001). Sensory and multisensory responses in the newborn monkey superior colliculus. *Journal of Neuroscience*, *21*(22), 8886–8894.

Wallace, M. T., & Stein, B. E. (2007). Early experience determines how the senses will interact. *Journal of Neurophysiology*, *97*(1), 921–926.

Ward, J., Huckstep, B., & Tsakanikos, E. (2006). Sound-colour synaesthesia: To what extent does it use cross-modal mechanisms common to us all? *Cortex*, *42*(2), 264–280.

Warren, W. H., & Verbrugge, R. R. (1984). Auditory perception of breaking and bouncing events: a case study in ecological acoustics. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(5), 704.

Wenzel, E. M., Arruda, M., Kistler, D. J., & Wightman, F. L. (1993). Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America*, *94*(1), 111–123.

Westbury, C. (2005). Implicit sound symbolism in lexical access: Evidence from an interference task. *Brain and Language*, *93*(1), 10–19.

Wozny, D. R., & Shams, L. (2011). Recalibration of auditory space following milliseconds of cross-modal discrepancy. *Journal of Neuroscience*, *31*(12), 4607–4612.

Xu, J., Yu, L., Rowland, B. A., Stanford, T. R., & Stein, B. E. (2012). Incorporating cross-modal statistics in the development and maintenance of multisensory integration. *Journal of Neuroscience*, *32*(7), 2287–2298.

Yuval-Greenberg, S., & Deouell, L. Y. (2009). The dog's meow: asymmetrical interaction in cross-modal object recognition. *Experimental Brain Research*, *193*(4), 603–614.

Zahorik, P. (2002). Auditory display of sound source distance. In *Proc. Int. Conf. on Auditory Display* (pp. 326–332).

Zampini, M., Guest, S., Shore, D. I., & Spence, C. (2005). Audio-visual simultaneity judgments. *Perception & Psychophysics*, *67*(3), 531–544.

Zmigrod, S., & Hommel, B. (2010). Temporal dynamics of unimodal and multimodal feature binding. *Attention, Perception, & Psychophysics*, *72*(1), 142–152.

Zmigrod, S., & Zmigrod, L. (2015). Zapping the gap: Reducing the multisensory temporal binding window by means of transcranial direct current stimulation (tDCS). *Consciousness and Cognition*, *35*, 143–149.